BAB I

PENDAHULUAN

1.1 Latar Belakang

Bahasa Indonesia merupakan bahasa nasional yang digunakan oleh masyarakat yang berasal dari berbagai latar belakang etnis, budaya dan wilayah geografis. Keberagaman tersebut melahirkan variasi aksen daerah dalam pengucapan Bahasa Indonesia, seperti aksen Jawa, Sunda, Batak, Bali dan berbagai aksen daerah lainnya. Variasi aksen tersebut tercermin dalam perbedaan pelafalan fonem, intonasi, ritme bicara, serta karakteristik akustik lainnya. Dalam komunikasi sehari-hari, perbedaan aksen dapat menimbulkan hambatan pemahaman, terutama pada interaksi lintas daerah, dan menjadi tantangan tersendiri dalam pengembangan teknologi pengenalan suara otomatis.

Bahasa Indonesia, sebagai bahasa resmi negara dan bahasa pengantar di antara berbagai suku di Indonesia, memiliki beragam aksen daerah yang mempengaruhi cara pengucapan dan pemahaman kata. Untuk mengatasi permasalahan ini, penting untuk mengembangkan suatu sistem yang mampu mendeteksi dan mengidentifikasi aksen-aksen tersebut secara otomatis. Salah satu teknologi yang menjanjikan dalam hal ini adalah Convolutional Neural Networks (CNN) dan teknik augmentasi audio, yang dapat meningkatkan akurasi dalam pengenalan aksen yang bervariasi. Penelitian terdahulu menunjukkan bahwa CNN dapat diaplikasikan dalam berbagai domain pengenalan suara, termasuk untuk bahasa daerah seperti Jawa menggunakan metode pengolahan sinyal dan citra yang tepat(Rahmawati dkk., 2021).

Implementasi CNN dalam deteksi aksen daerah dapat memberikan wawasan baru mengenai kompleksitas aksen yang terdapat dalam bahasa Indonesia. Mengingat kepadatan variasi aksen yang terdapat dalam masyarakat, hasil dari penelitian menunjukkan bahwa sistem dengan pendekatan CNN menunjukkan performa yang baik dalam klasifikasi aksen, dengan tingkat akurasi yang meningkat setelah penerapan augmentasi audio(Ripera, 2024). Namun, tantangan yang tersisa adalah pengumpulan data yang cukup representatif dari tiap aksen, mengingat kondisi demografi dan kebudayaan yang beragam di tiap daerah.

Berdasarkan penerapan machine learning terutama CNN dalam pengenalan bahasa dan aksen menunjukkan hasil yang menjanjikan, tetapi sangat tergantung pada kualitas dan keragaman data yang digunakan dalam pelatihan. Penelitian ini menyoroti pentingnya eksplorasi lebih lanjut dalam pengembangan dataset yang mencakup berbagai variasi aksen untuk mencapai sistem yang lebih komprehensif. (Andika dkk., 2023).

Automatic Speech Recognition (ASR) adalah teknologi yang memungkinkan sistem *computer* untuk mengenali, memproses, dan mengonversi suara manusia menjadi bentuk teks secara otomatis. Teknologi ini telah banyak diterapkan dalam berbagai aplikasi, seperti asisten virtual, sistem layanan pelanggan otomatis, navigasi berbasis suara, serta aplikasi Pendidikan dan Kesehatan. Meskipun demikian, performa sistem ASR sangat dipengaruhi oleh variasi karakteristik penutur, termasuk perbedaan aksen dan dialek. Penelitian terbaru menunjukan bahwa sistem ASR cenderung mengalami penurunan akurasi Ketika dihadapkan pada data ucapan dengan aksen yang berbeda dari data latihnya. (Ahlawat dkk., 2025)

Variasi aksen dalam Bahasa Indonesia menyebabkan perbedaan pola akustik dan fonetik, sehingga sistem ASR membutuhkan model yang mampu mengenali karakteristik ini untuk meningkatkan akurasi. Sebagian besar riset ASR saat ini berfokus pada bahasa yang sumber dayanya melimpah seperti Bahasa inggris, sementara aksen lokal seringkali kurang mendapat perhatian karena keterbatasan data latih. Masalah kurangnya data ini dikenal sebagai *low-resource problem*, yang merupakan tantangan umum dalam pengembangan ASR yang mampu menangani variasi penutur dan dialek. (Endah, Suprapto, dkk., 2025)

Untuk mengatasi tantangan data yang terbatas, pendekatan *deep learning* telah menjadi pilihan utama dalam penelitian ASR modern. Salah satu arsitektur yang sering digunakan adalah Convolutional Neural Network (CNN), yang mampu mengekstraksi fitur relevan dari representasi audio seperti spektrogram atau *Mel-Frequency Cepstral Coefficients* (MFCC). CNN telah terbukti efektif dalam berbagai tugas pemrosesan suara, termasuk klasifikasi audio dan pengenalan pola yang kompleks dalam sinyal suara (misalnya pada konteks emosi ucapan) menggunakan data augmented untuk memperkaya variasi pelatihan. Selain itu, kemampuan CNN untuk mengekstraksi fitur spasial dari data audio memungkinkannya membedakan ciri akustik unik dari berbagai aksen. Beberapa penelitian telah menunjukkan bahwa CNN dapat digunakan untuk

mengklasifikasikan aksen dengan tingkat akurasi yang sangat tinggi. (Barhoumi & BenAyed, 2025)

Meskipun CNN menunjukan performa yang baik, model ini sangat bergantung pada jumlah dan kualitas data pelatihan. Ketika dataset terbatas, model CNN cenderung mengalami *overfitting* dan memiliki kemampuan generalisasi yang rendah. Teknik augmentasi audio menjadi solusi popular untuk memperluas variasi data pelatihan tanpa harus mengumpulkan data baru secara manual. Augmentasi audio bekerja dengan memodifikasi sinyal suara seperti melalui *time stretching*, *pitch shifting*, penambahan *background noise*, dan manipulasi spektral lainnya, sehingga model terpapar variasi yang lebih luas dari data suara. (Maskur & Zahra, 2025)

Selain memperkaya variasi data, augmentasi audio juga membantu model menjadi robust terhadap kondisi nyata seperti kebisingan latar, perbedaan kecepatan berbicara, serta aspek lain yang muncul di lingkungan penggunaan nyata. Penelitian yang fokus pada dialek atau *low-resource dialectal ASR* menunjukan bahwa augmentasi data dapat meningkatkan performa sistem pengenalan aksen secara signifikan ketika dikombinasikan dengan pendekatan modern berbasis Transformer atau model neural lainnya. (Endah, Suprapto, dkk., 2025)

Dalam konteks Indonesia sendiri, beberapa penelitian telah mulai meneliti aspek variasi aksen atau dialectal speech recognition, meskipun cakupannya masih terbatas. Contohnya, penelitian mengenai interaksi manusia-mesin yang mencoba mengenali berbagai dialek lokal Nusantara menunjukan bahwa deep learning dengan transfer learning dapat membantu mengenali variasi aksen lokal dengan akurasi yang lebih baik daripada metode tradisional. (Suherwin dkk., 2025) Selain itu, penelitian teknik augmentasi pada data audio dialek Indonesia memberikan peningkatan makna dalam pengembangan model ASR untuk Bahasa lokal. (Endah, Suprapto, dkk., 2025)

Tujuan penelitian ini adalah untuk mengembangkan model yang dapat mendeteksi aksen daerah dalam ucapan bahasa Indonesia dengan menggunakan CNN dan teknik augmentasi audio. Diperkirakan, meskipun data pelatihan yang tersedia terbatas, metode ini akan memungkinkan model. Dasar tantangan utama berupa keterbatasan data representative dan keragaman variasi bicara yang kompleks, penelitian aksen daerah Indonesia semakin relevan, tidak hanya untuk meningkatkan performa sistem ASR tetapi

juga meningkatkan inklusivitas teknologi suara bagi penutur dari berbagai wilayah. Pendekatan yang memadukan CNN dengan augmentasi audio diharapkan dapat menyediakan metode yang lebih adaptif terhadap variasi aksen local, serta memberikan kontribusi penelitian baru dalam domain *natural language processing* dan *speech processing* untuk Bahasa Indonesia.

1.2 Identifikasi Masalah

a. Keterbatasan jumlah dataset aksen daerah menyebabkan model deep learning konvensional seperti CNN sulit mencapai performa optimal..
b. Model CNN dengan augmentasi audio belum mampu meningkatkan akurasi secara signifikan, yang dibuktikan dengan hasil akurasi sebesar 21% pada dataset 300 sampel.
c. Augmentasi audio tidak selalu mampu menambah informasi linguistik yang relevan, melainkan hanya menambah variasi sinyal tanpa memperkaya karakteristik fonetik yang membedakan aksen.
d. Kemampuan generalisasi model CNN pada data terbatas masih rendah, sehingga model kesulitan membedakan pola aksen yang memiliki kemiripan fonetik.
e. Belum optimalnya pendekatan supervised learning dalam kondisi few-data, sehingga diperlukan metode alternatif yang lebih adaptif terhadap dataset kecil.

1.3 Rumusan Masalah

1. Bagaimana performa model Convolutional Neural Network (CNN) dengan augmentasi audio dalam mendeteksi aksen daerah pada dataset terbatas.

2. Bagaimana performa model CNN dengan pendekatan Fine-Tuning Few-Shot Learning dalam mendeteksi aksen daerah pada dataset yang sama.

3. Apakah pendekatan Fine-Tuning Few-Shot Learning mampu meningkatkan akurasi deteksi aksen daerah dibandingkan CNN dengan augmentasi audio.

4. Apakah peningkatan performa yang dihasilkan oleh metode Fine-Tuning Few-Shot Learning signifikan secara statistik dibandingkan metode CNN dengan augmentasi audio.

5. Metode manakah yang lebih optimal untuk diterapkan dalam sistem deteksi aksen daerah pada kondisi keterbatasan data.

1.4 Tujuan Penelitian

Penelitian ini memiliki beberapa tujuan dianataranya:

1. Tujuan utama dari penelitian ini adalah untuk mengembangkan dan memancarkan model deteksi aksen daerah dalam ucapan bahasa Indonesia menggunakan CNN, dengan teknik augmentasi audio dan pendekatan *fine tuning few-shot learning* .

2. Penelitian ini juga bertujuan untuk mengidentifikasi teknik augmentasi audio yang paling efektif dalam meningkatkan akurasi model, serta mengeksplorasi potensi penerapan model dalam berbagai aplikasi berbasis suara.

1.5 Batasan Penelitian

Penelitian ini berjalan secara terarah dan fokus pada pencapaian tujuan, maka ditetapkan batasan-batasan berikut:

1. Jenis Bahasa dan Aksen yang Diteliti, penaksenan meliputi beberapa bahasa dari pulau indonesia, seperti ada jawa barat, jawa timur, jawa tengah, jakarta, dan yogyakarta.

2. Jenis Data Ucapan, data yang digunakan adalah data sekunder yang berasal dari sebuah video berasal dari sosial media.

3. Model Pembelajaran Mendalam, Model dan arsitektur CNN.

4. Metode Transfer Learning dengan menggunakan Few-Shot

    a. Episodic Training

    b. Fokus Intonasi dan Ritme

5. Teknik Augmentasi, panel – panel pada augmentasi ini berupa :

    a. Pergeseran Nada

    b. Peregangan Waktu

    c. Menambahkan Latar Belakang

    d. Teknik Perturbasi Volume

1.6 Manfaat Penelitian

Beriku ini manfaat penelitian ini diperuntukkan untuk:

1. Sastra Tambahan, Hasil penelitian ini dapat menjadi referensi bagi penayang tertarik.

2. Penerapan Augmentasi Audio pada, Penelitian ini memperkaya literatur tentang efek.

3. Pengembangan Sistem, Model yang dikembangkan dapat membantu membangun *ASR (Automatic Speech Recognition)* yang lebih adaptif terhadap variasi aksen daerah.

4. Peluang Pengembangan Teknologi Multibahasa dan Multilokal, Dengan pendekatan yang dikembangkan, penelitian ini membuka jalan bagi pengembangan teknologi suara yang mampu mengenali dan memahami keragaman, mendorong hadirnya teknologi yang lebih lokal.