BAB I
PENDAHULUAN

1.1. Latar Belakang

Teknologi pengenalan suara telah berkembang dan semakin banyak penggunaannya secara praktis. Teknologi ini memproses sinyal suara menjadi sebuah urutan kata atau entitas linguistik melalui algoritma tertentu yang diterapkan pada sebuah perangkat atau komputer (Huang dkk., 2001). Seiring dengan pertumbuhan *big data* dan daya komputasi yang besar, teknologi pengenalan suara telah berkembang luas sehingga memungkinkan penerapannya di berbagai aplikasi berbasis suara. Sebagai contoh adalah pencarian suara Google, asisten digital dan interaksi dengan perangkat seluler Siri dan Google Now pada perangkat Apple dan Android, kontrol suara Bardi Home di sistem rumah pintar, sistem hiburan Apple Car Play di dalam kendaraan mobil, dan berbagai aplikasi lainnya yang memanfaatkan pemrosesan sinyal suara (Jinyu, Li dkk., 2016).

Aplikasi dari sistem pengenalan suara biasanya dibangun dan dilatih di sebuah laboratorium dengan suara yang bersih dan homogen. Namun, kinerja sistem yang dibangun dengan kondisi tersebut dapat menurun secara signifikan ketika diterapkan di dunia nyata karena keadaan saat proses pengambilan data suara yang digunakan di dalam laboratorium tidak cocok dengan keadaan pengucapan suara di dunia nyata (Huang dkk., 2001). Oleh karena itu, akurasi dan keandalan adalah ukuran utama pada keberhasilan sistem pengenalan suara.

Ada beberapa faktor penting yang dapat memengaruhi kinerja sistem pengenalan suara, yaitu variabilitas konteks, variabilitas gaya bicara, variabilitas pembicara, dan variabilitas lingkungan (Benzeghiba dkk., 2007; Huang dkk., 2001). Variabilitas konteks menunjukkan variasi makna kata dan komunikasi. Kata dengan pengucapan yang sama mungkin memiliki arti yang berbeda. Kemudian, variabilitas gaya bicara merupakan variasi dari tingkat kecepatan dan nada seseorang dalam mengucapkan kalimat. Lalu, variabilitas pembicara dapat didefinisikan sebagai orang yang mengucapkan sebuah kata atau kalimat. Ucapan yang dikeluarkan mencerminkan karakteristik fisik, usia, jenis kelamin, pita suara, dialek, dan lainnya. Dengan

demikian, pola bicara satu orang bisa sangat berbeda dari orang lain. Selain itu, variabilitas lingkungan juga menjadi faktor penting pada sistem pengenalan suara. Ketika seseorang berinteraksi dengan komputer melalui suara, terdapat kondisi di mana sinyal suara yang diterima oleh komputer tidak jelas atau tidak bersih dikarenakan suara yang bergema, penempatan mikrofon yang jauh dari sumber suara (*distant*), atau kebisingan acak pada latar belakang suara. Faktor variabilitas lingkungan di atas menjadi peran penting dalam penurunan kualitas suara (Das dkk., 2021).

Berbagai macam sumber suara yang menyatu di dalam data yang diakuisisi sering menjadi persoalan pada sistem pengenalan suara. Misalnya, kondisi percakapan di mana tidak hanya suara percakapan seseorang, tetapi juga ada gangguan yang berasal dari ruangan yang bergema dan mikrofon yang jauh dari sumber suara. Selain itu, lingkungan yang bising seperti suara aktivitas manusia, suara kendaraan, atau suara angin yang ikut serta dalam data yang diakuisisi juga mengganggu kinerja sistem pengenalan suara. Segala gangguan yang tidak diinginkan yang tercampur di dalam data sinyal suara disebut sebagai derau (Jinyu Li dkk., 2016). Suara yang tercampur dengan derau menyebabkan kerusakan dan distorsi pada sinyal suara sehingga kinerja sistem pengenalan suara akan menurun (Zixing Zhang dkk., 2018).

Berbagai macam metode telah dikembangkan para peneliti untuk mengatasi masalah di atas. Pengembangan yang dilakukan secara umum terbagi menjadi dua, yaitu pengembangan pada domain fitur dan domain model (Jinyu Li dkk., 2014). Pengembangan pada domain fitur diperlukan sebuah pemahaman khusus tentang pemrosesan sinyal digital dan keterampilan dalam membangun persamaan matematika. Sementara itu, pengembangan pada domain model difokuskan untuk mengeksplorasi struktur model *deep learning* berdasarkan data sinyal suara yang telah melalui proses ekstraksi ciri, sehingga lebih mudah diterapkan dan tidak membutuhkan eksplorasi persamaan matematika secara eksplisit.

Metode *deep learning* digunakan sebagai sistem pengenalan suara karena memiliki kemampuan untuk memodelkan hubungan yang kompleks antar fitur dari sinyal suara (Deng & Li, 2013). Algoritma *deep learning* seperti *Deep Neural Network* (DNN) dan *Convolutional Neural Network* (CNN) digunakan pada

penelitian yang dilakukan oleh (Xuejiao Li & Zhou, 2016) di mana metode ekstraksi ciri *mel-frequency cepstral coefficient* (MFCC) digunakan sebagai masukan pada model DNN dan CNN. Model CNN mengungguli model DNN dengan skor akurasi 94,5%, sementara metode DNN memperoleh skor akurasi 71,9%. Modifikasi ekstraksi ciri dengan mentransformasi MFCC telah dilakukan oleh Pardede, dkk. sehingga menghasilkan lebih banyak fitur diskriminatif yang menjadi masukan pada algoritma CNN. Dengan demikian sistem pengenalan suara dapat mengenali suara yang bergema dengan *word error rate* yang rendah senilai 30,04% (Pardede dkk., 2018). Metode CNN dengan menggunakan skema pembagian bobot terbatas telah dilakukan oleh Abdel-Hamid dkk. dan dapat memodelkan fitur suara dengan lebih baik sehingga *error* yang diperoleh menurun sekitar 6%-10% dibandingkan dengan metode DNN (Abdel-Hamid dkk., 2014)

Metode CNN yang digunakan sebagai model *deep learning* untuk sistem pengenalan suara memiliki sifat *pooling* yang digunakan untuk menentukan korelasi antar fitur pada dimensi sinyal masukan secara lokal (X. Li & Zhou, 2016). Hal tersebut juga didukung dengan metode ekstraksi ciri MFCC yang bisa direpresentasikan sebagai matriks. Sehingga, CNN dapat mempelajari fitur-fitur dengan menekankan informasi pada data sinyal (Pardede dkk., 2018).

Metode ekstraksi ciri MFCC biasanya dipilih sebagai metode ekstraksi ciri sinyal suara dan sebagai masukan pada algoritma *deep learning*. Proses pada ekstraksi ciri MFCC meliputi *windowing*, transformasi fourier diskrit (DFT), mel *filter-bank*, kompresi menggunakan fungsi logaritmik, dan transformasi kosinus diskrit (DCT). Pada prosesnya, penggunaan fungsi logaritmik pada MFCC mempunyai tujuan tertentu. Pertama, mengubah hubungan antara saluran vokal. Kedua, mengompresi dinamika pada sinyal suara. Ketiga, merepresentasikan persepsi manusia dalam hal pendengaran. Hal tersebut membuat MFCC memperoleh kinerja yang baik dalam kondisi sinyal suara bersih, namun tidak tahan terhadap sinyal suara berderau (Pardede, 2016). Fungsi logaritmik yang ada pada MFCC meningkatkan sensitivitas terhadap perubahan di daerah energi rendah di mana banyak informasi berada sehingga akan menyebabkan ketidakcocokan yang signifikan ketika daerah ini dirusak oleh derau (Pardede, 2016).

Metode untuk mengatasi kekurangan pada MFCC telah dilakukan dengan menggunakan *power-law* (fungsi pangkat) untuk menggantikan fungsi logaritmik. Penggunaan *power-law* pada proses ekstraksi ciri telah terbukti meningkatkan keandalan terhadap derau (Lim, 1979). Penggunaan nilai pangkat yang sesuai pada *power-law* yang digunakan pada proses ekstraksi ciri dapat memberikan kompresi yang lebih baik di daerah energi rendah sehingga tidak sensitif ketika sinyal suara terdistorsi oleh derau (Lockwood & Alexandre, 1994). Salah satu metode ekstraksi ciri yang menggunakan *power-law* adalah *Power-Normalized Cepstral Coefficient* (PNCC) (Kim & Stern, 2009, 2010, 2016). PNCC menggunakan *power-law* non-linear untuk menggantikan fungsi logaritmik pada MFCC sehingga dapat meredam derau berdasarkan tapis asimetris (Kim & Stern, 2016).

Metode ekstraksi ciri MFCC dan PNCC dilakukan melalui tahap DCT. Proses DCT pada metode ekstraksi ciri tersebut menghilangkan informasi berupa korelasi antar komponen fitur (Yuliani dkk., 2017). Oleh karena itu, penggunaan metode ekstraksi ciri primitif lebih menarik untuk digunakan untuk model yang sederhana seperti DNN (Pardede dkk., 2019). *Filter-bank* (FBANK) merupakan salah satu jenis metode ekstraksi ciri primitif. Berbeda dengan MFCC, FBANK tidak melewati tahap transformasi kosinus diskrit, melainkan hanya melewati tahap ekstraksi spektrogram melalui *short-time fourier transform* (STFT) dan penskalaan logaritmik terhadap setiap filter. FBANK lebih dapat diterima sebagai ekstraksi ciri terhadap DNN dikarenakan mengekstrak lebih banyak fitur sehingga tahan terhadap variasi yang rendah pada sinyal suara.

Berbagai metode ekstraksi ciri di atas menjadi masukan pada model *deep learning*. Namun, model *deep learning* yang banyak digunakan pada pengenalan suara hanya melibatkan model tunggal yang memiliki kemampuan belajar terbatas (Chu dkk., 2017). Penggabungan beberapa model *deep learning* dapat digunakan untuk mengatasi kelemahan model tunggal (Chu dkk., 2017; Yao dkk., 2018). Beberapa penelitian tentang pengenalan suara menggunakan gabungan model *deep learning* telah dilakukan. Gabungan empat algoritma *machine learning* yang digunakan oleh (Guo dkk., 2022) menunjukkan efektivitas stabilitas konstruksi fitur sehingga mendapatkan akurasi 94% untuk sistem pengenalan emosi suara. CNN

dan BLSTM yang disandingkan secara seri digunakan oleh (Naiborhu & Endah, 2021) untuk mengenali lima dialek suara Bahasa Indonesia yang berbeda. Gabungan algoritma *support vector machine* (SVM) dengan *deep neural network* (DNN) telah digunakan oleh Ariyanti, dkk. untuk mengenali suara patologis berdasarkan kombinasi antara sinyal akustik dengan rekam medik (Ariyanti dkk., 2021). Penelitian-penelitian tersebut menunjukkan bahwa menggabungkan beberapa model cenderung mengungguli model tunggal.

Berdasarkan pada masalah latar belakang di atas, maka dalam penelitian ini dilakukan pengembangan pada domain model dengan menggabungkan dua algoritma *deep learning* secara paralel untuk sistem pengenalan suara berbasis *hybrid deep learning* menggunakan ekstraksi ciri berbasis *power-law*. Melalui metode di tersebut diperoleh sistem pengenalan suara dengan akurasi tinggi dan andal terhadap derau.

1.2. Perumusan Masalah

Berdasarkan pada latar belakang di atas maka ada dua permasalahan yang diteliti di dalam penelitian ini.

1. Manakah dari dua metode ekstraksi ciri berbasis fungsi logaritmik dan berbasis *power-law* yang memberikan kinerja model klasifikasi yang paling optimal untuk sistem pengenalan perintah suara?
2. Apakah metode *hybrid deep learning* memberikan akurasi klasifikasi lebih tinggi dibanding metode *deep learning* tunggal pada sistem pengenalan perintah suara?

1.3. Tujuan Penelitian

Tujuan penelitian ini secara umum adalah untuk mengembangkan sistem pengenalan suara berbasis *deep learning*. Secara spesifik, tujuan penelitian adalah sebagai berikut:

1. Menentukan metode ekstraksi ciri yang paling efektif untuk klasifikasi pengenalan perintah suara.
2. Mengevaluasi kinerja model *deep learning* untuk klasifikasi pengenalan perintah suara.

1.4. Manfaat Penelitian

Beberapa manfaat yang dapat diperoleh dari hasil penelitian ini antara lain sebagai berikut:

1. Dari sisi pengembangan keilmuan penelitian ini diharapkan dapat memberikan kontribusi di bidang pengolahan suara dalam menentukan metode ekstraksi ciri dan metode klasifikasi yang tepat sehingga menghasilkan akurasi klasifikasi secara optimal;
2. Secara praktis penelitian ini diharapkan dapat memberikan alternatif solusi dalam pengembangan sistem pengenalan suara yang andal terhadap derau.