

BAB I PENDAHULUAN

1.1 Latar Belakang

Kemampuan sistem pendengaran manusia dalam mengenal suara dapat ditiru untuk diimplementasikan pada mesin, robot, sistem keamanan, dan lain-lain. Komponen- komponen pendukungnya telah tersedia, yaitu *microphone* sebagai pengganti telinga, algoritma- algoritma pemrosesan sinyal, dan Jaringan Saraf Tiruan (JST) yang menjalankan fungsi sebagai otak. JST digunakan karena memiliki kemampuan untuk mengenali pola yang sedikit berubah dari aslinya. Hal ini mirip dengan otak manusia yang masih mampu untuk mengenali seseorang dari suara yang sedikit berubah misalnya lewat telepon. JST merupakan model sistem komputasi yang bekerja seperti sistem saraf biologis. JST berfungsi sebagai pengganti saraf dan otak, dimana pada saat berhubungan dengan dunia luar, kemampuan belajar dan generalisasinya cepat dan mudah dalam pengenalan pola suatu karakter serta mudah untuk diimplementasikan. Sistem pendengaran tiruan tersebut dapat dirancang sedemikian rupa dengan meniru sebagian prinsip kerja sistem pendengaran manusia sehingga nantinya sistem tersebut dapat mengenali suara berdasarkan gelombang suara yang diterimanya [1].

Suara yang didengar sehari-hari adalah merupakan gelombang analog. Gelombang ini berasal dari tekanan udara yang ada di sekitar, yang dapat didengar dengan bantuan gendang telinga. Gendang telinga ini bergetar, dan getaran ini dikirim kemudian diterjemahkan menjadi informasi suara yang dikirimkan ke otak, sehingga dapat mendengarkan suara. Namun, komputer hanya mampu mengenali suara dalam bentuk sinyal digital. Sehingga diperlukannya proses mengubah sinyal analog menjadi sinyal digital. Pengolahan sinyal merupakan hal penting dalam ilmu pengetahuan dan teknologi komunikasi. Salah satu bidang pengolahan sinyal yang berpengaruh dalam teknologi komunikasi adalah *speech recognition* [2].

Speech recognition adalah salah satu cara untuk memudahkan bagi siapa saja untuk mengoperasikan komputer terutama bagi orang-orang dengan keterbatasan

pendengaran. *Speech recognition* menjadi faktor penting dari interaksi antara manusia dan komputer [3]. Teknologi *speech recognition* dikembangkan untuk mewujudkan keinginan manusia dalam memaksimalkan fungsi komputer sebagai alat yang mampu mempermudah pekerjaan manusia [4]. Penelitian ini difokuskan untuk dapat mengenali angka (*zero, one, two, three, four, five, six, seven, eight, dan nine*). Angka tersebut memiliki cara pengucapan yang berbeda sehingga menghasilkan ciri sinyal yang berbeda pula [5].

Terdapat penelitian terkait tentang *digit speech recognition*, diantaranya yaitu pada tahun 2010, Ghulam Muhammad dkk [6], melakukan penelitian tentang *speech recognition* digit 0 sampai 9 dalam Bahasa Bangla menggunakan metode *Mel-Frequency Cepstral Coefficients* (MFCC) untuk klasifikasi fitur dan *Hidden Markov Model* untuk pengenalan. Hasil yang didapatkan dari penerapan dua metode tersebut tingkat keakuratan masih rendah dan komputasi lama. Metode MFCC dinilai masih memiliki keterbatasan untuk mengklasifikasikan pengenalan suara dengan data yang besar.

Penelitian selanjutnya menggunakan *Convolutional Neural Network* (CNN) telah terbukti mengungguli jaringan hybrid *Deep Neural Network / Hidden Markov Model* (DNN/HMM) untuk *speech recognition*. Pada tahun 2014, Ossama Abdel-Hamid dkk [7], melakukan penelitian tentang *TIMIT phone recognition and the voice search large vocabulary speech recognition tasks* dengan metode CNN. Penelitian tersebut menunjukkan bahwa CNN mengurangi tingkat *error* sebesar 6%-10% dibandingkan dengan DNN.

Penelitian terkait terakhir pada tahun 2019, Shuvo dkk [8], melakukan penelitian dengan memanfaatkan MFCC sebagai ekstraksi fitur dan CNN untuk klasifikasi fitur yang berjudul *Bangla Numeral Recognition from Speech Signal Using Convolutional Neural Network*. Sistem yang diusulkan dinilai mengungguli sebagian besar penelitian serupa dengan memperoleh akurasi pengenalan sebesar 93,65%.

Oleh karena itu penelitian ini menggunakan metode CNN yang menggabungkan proses ekstraksi dan klasifikasi fitur menjadi satu *network* [9][10]. CNN dirancang khusus untuk pengenalan serta klasifikasi gambar dan

video. Kemudian pengenalan dilakukan dengan menginputkan suara langsung ke sistem. Selanjutnya sistem menghasilkan *output text* yaitu "zero", "one", hingga "nine". CNN memiliki beberapa lapisan (*layer*) yang mengekstrak informasi dari gambar dan menentukan klasifikasi dari gambar tersebut [11].

Dalam penelitian ini membangun sistem untuk *speech recognition* yang berfokus pada angka 0 sampai 9. Dataset sebagai masukan untuk pelatihan dan pengujian. Dataset *speech commands* berjumlah 23.666 rekaman dengan 10 kelas. Dengan menerapkan metode CNN sebagai salah satu cara untuk *digit speech recognition*.

1.2 Perumusan Masalah

1.2.1 Permasalahan

MFCC dan HMM memiliki keterbatasan dalam melakukan *speech recognition multiclass* dengan data yang besar karena harus menggunakan model untuk proses klasifikasi.

1.2.2 Metode Usulan

Penelitian ini menggunakan metode *Convolutional Neural Network* (CNN) yang umumnya memiliki dua bagian utama yaitu ekstraksi dan klasifikasi fitur yang berfungsi untuk mengoptimalkan komputasi serta mendapatkan tingkat keakuratan yang lebih baik dalam *speech recognition*.

1.2.3 Pertanyaan Penelitian

Pertanyaan pada penelitian ini sebagai berikut :

1. Bagaimana penerapan metode CNN dengan spektogram mampu melakukan *recognition* suara manusia?
2. Berapa nilai *precision*, *recall* dan akurasi metode CNN yang digunakan?
3. Berapa nilai akurasi *recognition* menggunakan *dataset* lain?

1.3 Tujuan & Manfaat

1.3.1 Tujuan Penelitian

Tujuan dalam penelitian ini adalah mengimplementasikan *deep learning* dengan metode CNN pada *recognition* suara manusia (*zero, one, two, three, four, five, six, seven, eight, nine*) untuk mengetahui nilai *precision*, *recall* dan akurasi.

1.3.2 Manfaat

Manfaat pada penelitian ini sebagai berikut :

1. Sebagai pengembangan ilmu diketahui bahwa kemampuan metode CNN dalam melakukan *recognition* suara dengan baik.
2. Bagi pengembangan sistem selanjutnya dalam penelitian ini dapat digunakan sebagai dasar untuk mengembangkan sistem absensi, keamanan, entri pin dan lain-lain.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini sebagai berikut :

1. Data yang digunakan pada penelitian ini yaitu dataset *speech commands* di situs (https://storage.googleapis.com/download.tensorflow.org/data/speech_commands_v0.01.tar.gz) (online 17 Februari 2020).
2. *Dataset* berjumlah 23.666 rekaman berformat .wav dengan 10 kelas (*zero, one, two, three, four, five, six, seven, eight, nine*) yang digunakan pada penelitian ini.
3. *Dataset* untuk *recognition* yang digunakan pada penelitian ini yaitu *Spoken Digit Dataset* dari *kaggle* berjumlah 100 data.
4. Dalam skenario ujicoba , pencarian nilai akurasi terbaik dilakukan dengan cara inisialisasi parameter *training options* yaitu *learning rate, max epoch, validation frequency* dan *minibatch-size*. Dalam setiap ujicoba disesuaikan dengan *tools* yang digunakan.

1.5 Sistematika Proposal

Sistematika penulisan laporan tugas akhir ini adalah sebagai berikut :

1. BAB I PENDAHULUAN

Bab ini membahas tentang latar belakang mengenai ruang lingkup masalah yang dibahas, solusi yang diusulkan untuk dapat menyelesaikan permasalahan, tujuan dan manfaat dengan diterapkannya solusi yang diusulkan, serta sistematika penulisan laporan.

2. BAB II KAJIAN PUSTAKA

Pada bab ini dibahas mengenai metode usulan yang dipilih, contoh penggunaan metode dan penelitian terkait.

3. BAB III METODE USULAN

Pada bab ini dibahas mengenai tahapan penelitian, penjelasan arsitektur sistem, dan *dataset* yang digunakan.

4. BAB IV HASIL DAN PEMBAHASAN

Bab ini menguraikan implementasi dari metode penelitian pada bab sebelumnya, scenario pengujian, proses hasil ujicoba sistem, dan analisis hasil ujicoba sistem.

5. BAB V PENUTUP

Bab ini berisi kesimpulan dari penelitian yang disesuaikan dengan pertanyaan penelitian dan saran terkait pengembangan aplikasi selanjutnya.

6. REFERENSI

Berisi tentang daftar jurnal-jurnal atau referensi yang dijadikan acuan dalam penelitian ini.