EEICT
https://ojs.uniska-bjm.ac.id/index.php/eeict

e-ISSN: 2615-2169
p-ISSN: 2654-4296

ENGLISH SPOKEN DIGIT RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK (CNN)

Muhammad Aminuddin[1], Muhammad Rizqi Subeno[2]

[1,2]Prodi Magister Teknik elektro, Departemen Teknik Elektro dan Teknologi Informasi, Fakultas Teknik, Universitas Gadjah Mada
Email: muhammadaminuddin@mail.ugm.ac.id[1]

Received 05 July 2023 | Revised 26 Oct 2023 | Accepted 27 Oct 2023
DOI : 10.31602/eeict.v6i2.11877

EEICT (Electric, Electronic, Instrumentation, Control, Telecommunication)
disebar luaskan oleh : Lisensi Creative Commons Atribusi 4.0 Internasional.

Abstrak - *Digit Spoken recognition* atau pengenalan suara merupakan proses untuk menerjemahkan kata-kata menjadi sebuah teks (angka). Data yang diterjemahkan dapat berupa rekaman audio suara maupun menggunakan suara langsung. Pengenalan suara dapat membuat interaksi antara manusia kepada komputer menjadi lebih intuitif, aksebilitas bagi penyandang disabilitas, serta memungkinkan pengoperasian dengan *hands-free*. Dataset terdiri dari rekaman suara 6 *speaker* dalam bahasa inggris dengan total 3.000 data, yang terbagi sebanyak 80% untuk data latih dan 20% untuk data validasi. *Mel-frequency cepstrum coefficient* (MFCC) digunakan untuk mengekstraksi fitur dari rekaman audio suara. Selanjutnya data tersebut digunakan untuk melatih model CNN. Hasil dari model CNN yang diusulkan dapat mengklasifikasi data suara audio dengan akurasi lebih dari 90%.

Kata Kunci : CNN, Spoken, Digit, Recognition, Deep Learning

I. PENDAHULUAN

Speech recognition dalam bahasa indonesia adalah pengenalan suara yang merupakan suatu proses menerjemahkan kata-kata menjadi sebuah teks. *Speech recognition* melibatkan menangkap dan mengubah gelombang suara ke dalam bentuk digital mengubahnya menjadi bahasa dasar atau fonem serta menganalisis kata-kata secara konstektual [1].

Deep Learning secara umum telah diaplikasi pada berbagai bidang seperti robotik, otomasi, analisis, peramalan, termasuk dalam bidang pengenalan suara. Terdapat metode-metode pada *deep learning* yang berfungsi untuk mengidentifikasi atau mengklasifikasi, yang mana hal tersebut sesuai untuk diimplentasikan pada *speech recognition*. *Deep learning* yang digunakan untuk *automated speech recognition* (ASR) telah banyak dipelajari secara ekstensif. Sistem ASR didasarkan pada kemampuan untuk mengenali kalimat, kata, suku kata dan fenom. Arsitektur *deep learning* yang

e-ISSN: 2615-2169
p-ISSN: 2654-4296

umum digunakan adalah *recurrent neural network* (RNN) dan *convolutional neural network* (CNN) [2], [3].

Pada awalnya arsitektur CNN ini diaplikasikan pada masalah *computer vision*. Selanjutnya, CNN juga diaplikasikan pada bidang pengenalan suara dan berhasil untuk mengidentifikasi atau mengklasifikasi, baik menggunakan arsitektur CNN 1-D, CNN 2-D dan CNN 3-D. Sebagian besar fitur ucapan yang digunakan untuk pelatihan pada CNN didasarkan pada *fast fourier transporm* (FFT), Mel Frequency Cepstral coefficient (MFCC), Mel Filter Bank (MFB), dan *spectrum envelopes* [4], [5].

Kebutuhan akan pengenalan suara berawal dari keinginan untuk dapat membuat interaksi yang lebih alami dan intuitif kepada komputer. Selain itu pengenalan suara juga dapat digunakan untuk aksebilitas bagi penyandang disabilitas serta memungkinkan pengoperasian perangkat *hands-free* [1]. Pada paper ini, kami mengusulkan metode deep learning, yaitu CNN untuk dapat mengenali dan mengklasifikasikan file audio suara menjadi informasi digit 0-9.

II. METODE

1. Dataset

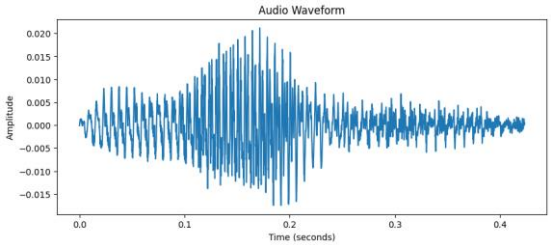
Dataset berupa file audio suara dengan format .wav yang terdiri dari suara 6 orang berbeda mengucapkan digit 0-9 dalam bahasa inggris. Setiap orang memiliki rekaman sebanyak 50 file audio pada setiap digit, sehingga total dataset yang dimiliki sebanyak 3.000 file audio. Dataset ini nantinya akan dibagi menjadi 80% untuk data uji dan 20% untuk data validasi.

Tabel 1. Dataset untuk pelatihan CNN

Nama *Speaker*	Jumlah rekaman audio file
George	500
Jackson	500
Lucas	500
Nicolas	500
Theo	500
Yweweler	500
Total	3.000

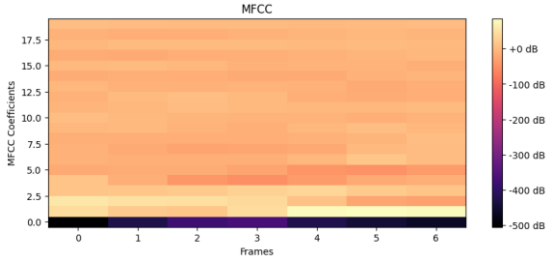
2. *Ekstraksi Fitur*

Ekstraksi fitur bertujuan mengkonversi file audio yang berupa gelombang suara menjadi M*el Frequency Cepstral Coeffiecient* (MFCC) untuk mendapatkan fitur dari dataset yang dimiliki. Kemudian dengan menggunakan modul librosa, fitur MFCC diekstraksi dari rekaman suara angka dan disimpan dalam bentuk *array*. Adapun gambar dari gelombang suara dan MFCC dari file audio dataset adalah sebagai berikut.



Gambar 1. Gelombang suara angka-0

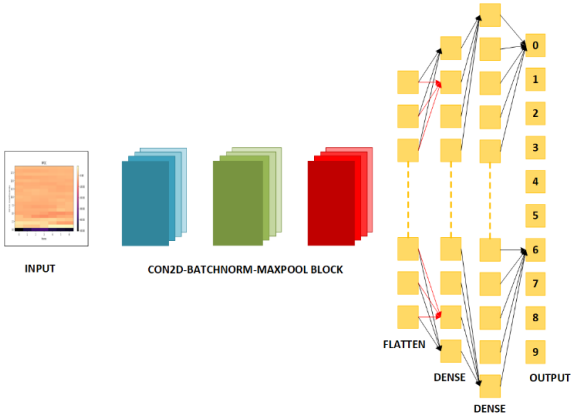
e-ISSN: 2615-2169
p-ISSN: 2654-4296



Gambar 2. MFCC suara angka-0

3. Arsitektur CNN

Arsitektur CNN yang digunakan dapat dilihat pada gambar 3. Bentuk input model CNN adalah (20*20*1). Tahapan pertama pada pelatihan adalah melakukan konvolusional 2-D dengan ukuran filter 32 untuk konvolusi pertama, 48 filter untuk konvolusi kedua dan 120 filter untuk konvolusi ketiga. Ukuran kernel adalah (2,2) dengan aktivasi relu, kemudian dinormalisasi. Lapisan *batchnormalization* terakhir (lapisan ketiga berwarna merah) kemudian ditransfer ke lapisan *maxpooling* yang memiliki ukuran *pool size* (2, 2).

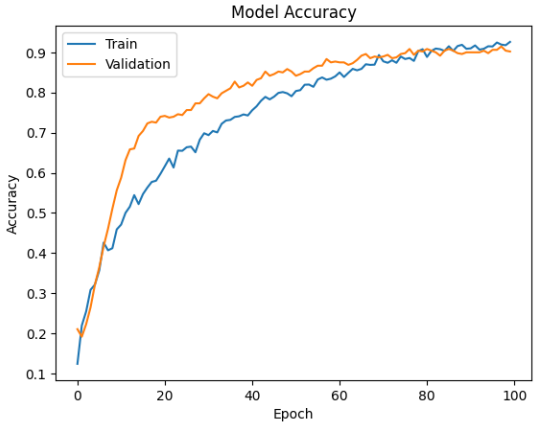


Gambar 3. Arsitektur CNN yang diusulkan

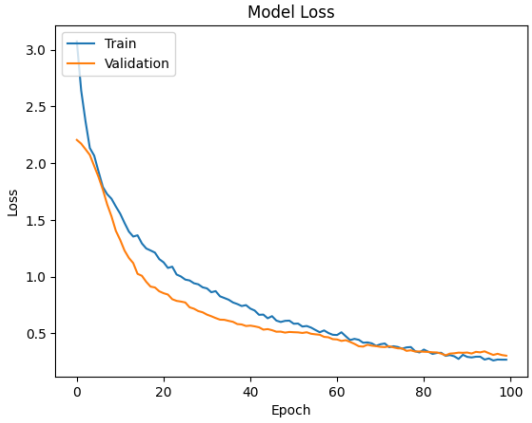
Selanjutnya keluaran dari lapisan *maxpooling* diubah ke vektor 1-D seperti pada bagian *flatten*, dan kemudian saling terhubung dengan layer-layer dense. Bagian ini sama seperti pelatihan jaringan *neural network*. Aktivasi yang digunakan adalah relu pada tiap bagian. Output yang diharapkan adalah klasifikasi nilai digit 0-9, maka dari itu untuk dapat melakukan klasifikasi tersebut digunakan fungsi softmax.

III. HASIL DAN PEMBAHASAN

Dari berbagai eksperimen terhadap pelatihan jaringan CNN, didapatkan hasil pada gambar 4 dan gambar 5 berikut.



Gambar 4. Akurasi model CNN yang diusulkan



Gambar 5. *Loss* model CNN yang diusulkan

e-ISSN: 2615-2169
p-ISSN: 2654-4296

Model CNN yang diusulkan diberikan pelatihan sebanyak 100 epochs, dengan dioptimisasi menggunakan Adadelta optimizer. Akurasi dari model yang didapat adalah lebih dari 90%, yang mana ini menyatakan bahwa model yang didapat dari hasil pelatihan cukup akurat. Selanjut dengan dataset untuk validasi dilakukan pengujian model pada tiap kelas untuk mengetahui keakuratan model dalam mengklasifikasi tiap kelas. Adapun hasilnya ditunjukan pada tabel 2 berikut.

Tabel 2. Pengujian akurasi tiap kelas

Class	*Accuracy*
Class 0	1.0
Class 1	0.74
Class 2	0.95
Class 3	0.81
Class 4	0.89
Class 5	0.93
Class 6	0.92
Class 7	1.0
Class 8	0.95
Class 9	0.9

Dari tabel 2 dapat terlihat ketika dilakukan pengujian pada model yang telah dilatih secara umum akurasi pada tiap kelas berada diatas 0.8, meskpiun pada kelas 1 akurasi yang didapat masih 0.74, secara keseluruh model CNN tersebut dapat dikatakan baik. Metriks evaluasi model dirangkum pada tabel 3 sebagai berikut.

Tabel 2. Metriks evaluasi model CNN

Class	*Precision*	*Recall*	*F1-soce*	*Support*
Class 0	0.98	1.00	0.99	42
Class 1	0.88	0.74	0.80	57
Class 2	0.97	0.95	0.96	60
Class 3	0.93	0.81	0.87	53
Class 4	1.00	0.89	0.94	53
Class 5	0.88	0.93	0.90	40
Class 6	0.88	0.92	0.90	38
Class 7	0.84	1.00	0.92	49
Class 8	0.90	0.95	0.92	38
Class 9	0.79	0.90	0.84	50

Precision menunjukan ukuran seberapa banyak prediksi positif yang benar dari semua prediksi positif yang dilakukan untuk suatu kelas. Recall menunjukan ukuran seberapa banyak prediksi positif yang benar dari semua contoh di kelas tersebut. F1-score adalah ukuran gabungan yang menggabungkan precission dan recall dalam satu angka. Sedangkan support adalah sampel dalam setiap kelas yang digunakan dalam pengujian.

IV. KESIMPULAN

1. Model arsitektur deep learning, yaitu CNN dapat digunakan untuk mengklasifikasikan data suara digit dalam bahasa inggris.
2. Kompleksitas model arsitektur CNN, jumlah dan pengolahan dataset, serta banyaknya epochs dalam pembelajaran model akan berpengaruh terhadap akurasi model CNN.

DAFTAR PUSTAKA

[1] Alfiandra. 2009. Kajian partisipasi masyarakat yang melakukan pengelolaan persampahan 3R di Kelurahan Ngaliyan, Kalipancur Kota Semarang. Tesis : Universitas Diponegoro.

[2] Asip, F., T. anggun dan N. Fitri. 2014. Pembuatan Briket Dari

e-ISSN: 2615-2169
p-ISSN: 2654-4296

Campuran Limbah Plastik LDPE, Tempurung Kelapa Dan Cangkang Sawit. *Jurnal Teknik Kimia.* 2 (20): 46-48.

[3] Faizal, M. A. D. Rifky dan I. Sanjaya. 2018. Pembuatan Briket drai Campuran Limbah Plastik LSPE dan Kulit Buah Kapuk Sebagai Energi Lternatif. *Jurnal Teknik Kimia.* 1 (24): 9.

[4] Galgani, F., D. Fleet, J.V. Franeker, S. Katsanevakis, T. Maes, J. Mouat, L. Oosterbaan, I. Poitou, G. Hanke, R. Thompson, E. Amato, A. Birkun, dan C. Janssen. 2010. Marine Strategy Framework Directive – Task Group 10 Report Marine Litter. European Commission Joint Research Centre. Ispra. Available on JRC Publications Repository - Marine Strategy Framework Directive - Task Group 10 Report Marine Litter (europa.eu)

[5] Herliati, S. B. Prasetyo dan Y. Verinaldy. 2019. Potensi Limbah Plastik dan Biomassa Sebagai Sumber Energi Terbarukan Dengan Proses Pirolisis. Jurnal Teknologi. Vol. 6. Edisi 2. Hlm. 87-88.