Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 4. May 2024

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Penerapan Metode MFCC dan LSTM untuk Speech Emotion Recognition

I Dewa Agung Adwitya Prawangsa, AAIN Eka Karyawati

Informatika, Universitas Udayana
Jl. Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
wawaca.waca@gmail.com
eka.karyawati@unud.ac.id

Abstract

Speech recognition falls under the field of computational linguistics. This includes identification, recognition, and translation of speech detected into text by a computer. This research uses Mel Frequency Cepstral Coefficients (MFCC) and Recurrent Neural Networks (RNN) techniques as a form of Artificial Neural Network (ANN) architecture. The main objective of this research is to use speech recognition techniques to detect and identify various emotional voices within a person. The MFCC process will convert the voice signal into several vectors that help for the speech recognition process in this study. The results obtained are by determining two comparison parameters, namely a learning rate of 0.1 which results in an accuracy validation of 72% and a learning rate of 0.001 which results in 57%.

Keywords: *Feature Extraction**,** Speech recognition , MFFCC, ANN, RNN*

1. Pendahuluan

Emosi merupakan suatu kondisi mental seseorang yang dapat mendorongnya untuk melakukan suatu tindakan atau berekspresi yang dapat dipicu dari dalam atau luar dirinya. Dalam kehidupan sehari-hari sangat penting untuk memahami kondisi emosional seseorang dengan emosi tertentu. Emosi juga merupakan salah satu aspek penting bagi kehidupan [1]. Emosi seseorang dapat diketahui salah satunya dari ekspresi wajah, namun terkadang ekspresi wajah seseorang tidak sesuai dengan apa yang sedang dialaminya. Pada penelitian kali ini, Saya akan membuat aplikasi guna sebagai penelitian yang dapat mengetahui emosi yang sedang dialami oleh seseorang, dibuatlah aplikasi untuk mendeteksi emosi seseorang berdasarkan suara menggunakan metode Mel Frequency Cepstral Coefficients (MFCC) dan Long short term memory network (LSTM).

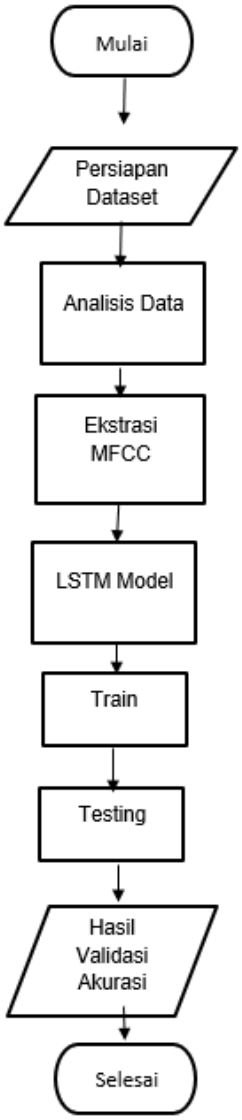
Penelitian ini juga berdasarkan dari beberapa penilitian terdahulu yang terkait yaitu pada jurnal pertama dari penulis Fatan Kasyidi, Ridwan Ilyas, Nida Muthi Annisa, dengan judul "Peningkatan Kemampuan Pengenalan Emosi melalui Suara dalam Bahasa Indonesia" Dalam review singkat pada jurnal ini berisikan tentang permasalahan tentang pengenalan emosi melalui suara erutama masalah korpus yang menjadi salah satu faktor yang menjadikan pengenalan emosi ini belum menghasilkan akurasi pengenalan yang optimal, khususnya berkaitan dengan imbalance data. Penelitian ini dilakukan untuk meningkatkan performa pengenalan emosi untuk mengenali lima kelas emosi yaitu senang, marah, sedih dan kepuasan serta netral menggunakan algoritma boosting. Selain itu, digunakan pula metode seperti CNN dan RNN untuk dapat dilakukan perbandingan serta penerapan SMOTE untuk korpusnya. Setelah eksperimen, dapat dihasilkan akurasi pengenalan mencapai 65% untuk akurasi untuk data tes berdasarkan konfigurasi 22050 Hz sebagai sampling rate, MFCCs dan oversampling SMOTE [1].

Lalu pada review jurnal penilitan yang kedua dengan penulisnya yang bernama Barlian Henryranu Prasetio, Wijaya Kurniawan, Mochammad Hannats Hanafi Ichsan, yang berjudul " Pengenalan Emosi Berdasarkan Suara Menggunakan Algoritma HMM" dijelaskan bahwa penilitian ini memiliki tujuan untuk mengenali emosi seseorang melalui ucapan menggunakan algoritma HMM. Sistem dibangun dapat mengenali 3 jenis emosi yaitu marah, bahagia dan netral. Fitur yang digunakan dalam sistem ini adalah

pitch, energi dan formant. Database yang digunakan adalah suara dari rekaman film. Dari hasil obeservasi probabilitas emosi marah sebesar 0.196, bahagia 0.254 dan netral 0.045. Sistem memiliki tingkat akurasi rata-rata sebesar 86.66%. Rata waktu eksekusi sistem dalam mendeteksi dan mengklasifikasikan emosi sebesar 21.6ms [3].

2. Metodolgi Penelitian

Flowchart Umum



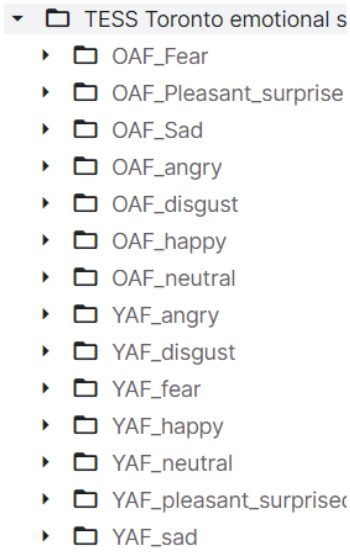
Gambar 1. Flowchart Alur Kerja

2.1 Persiapan Data Set

Mempersiapkan Data set yang diperlukan guna sebagai sample suara dalam melakukan pengenalan emosi dalam penelitian kali ini yaitu meggunakan Toronto emotional speech set (TESS) yang bersifat publik yang memudahkan dalam menggunakannya sebagai sample penlitian ini. Sumber dataset

p-ISSN: 2301-5373
e-ISSN: 2654-5101

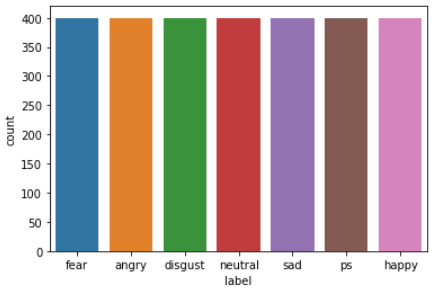
berapa pada tautan berikut ini : www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess. Ada satu set 200 kata target yang diucapkan dalam frasa pembawa "Ucapkan kata _' oleh dua aktris (berusia 26 dan 64 tahun) dan rekaman dibuat dari set yang menggambarkan masing-masing dari tujuh emosi (marah, jijik, takut, bahagia, kejutan, kesedihan, dan netral) Total ada 2800 titik data (file audio). Dataset diatur sedemikian rupa sehingga masing-masing dari dua aktor wanita dan emosi mereka terkandung dalam foldernya sendiri. Dan di dalamnya, semua file audio 200 kata target dapat ditemukan. Format file audio adalah format WAV.



Gambar 2. Toronto emotional speech set (TESS)

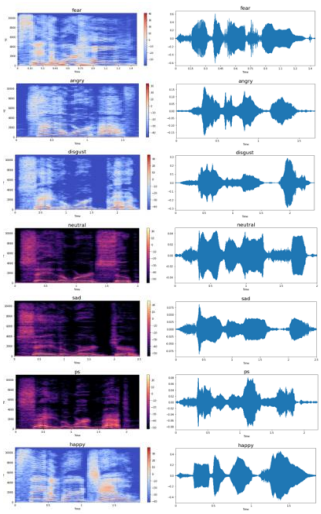
2.2 Analisis data

Dalam proses membuat data analisis, Saya akan membuat sebuah visual analisis yang pertama yaitu yang memuat label dari semua kelas dalam distribusi yang sama.



Gambar 3. Label semua kelas

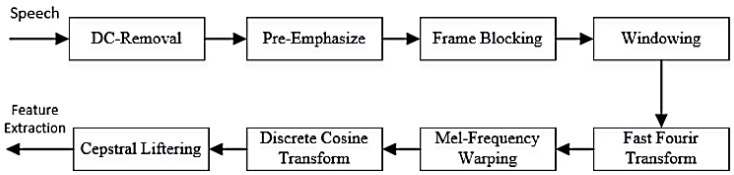
Lalu analisis yang kedua yaitu menampilkan waveform, spektogram serta menampilkan plot dari sample berkas yang digunakan.



Gambar 4. Spektogram & Waveform

2.3 Ekstrasi Fitur MFCC

Teknik Mel Frequency Cepstral Coefficients (MFCC) digunakan untuk ekstraksi ciri dari sinyal wicara dan membandingkan dengan penutur tak dikenal dengan penutur yang ada dalam database [2].



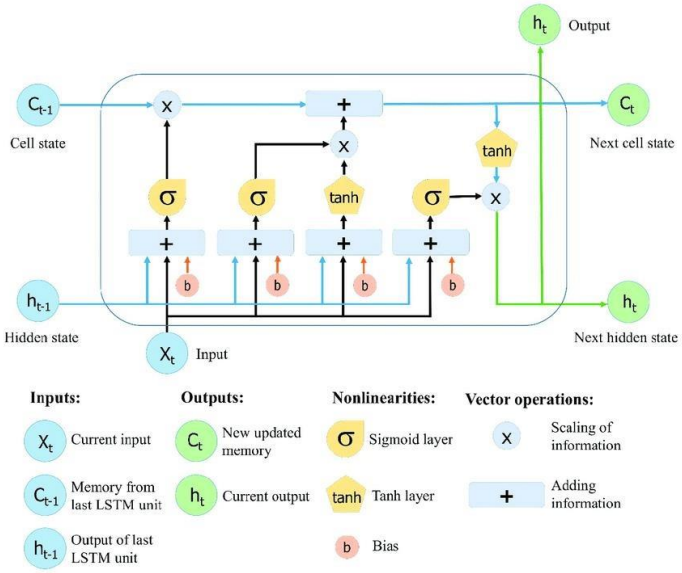
Gambar 5. Tahap proses Ekstrasi MFCC [6].

Durasi audio dibatasi hingga maksimal 3 detik untuk durasi ukuran file yang sama dan akan mengekstrak fitur koefisien cepstral frekuensi Mel (MFCC) dengan batas 40 dan mengambil rata-rata sebagai fitur akhir yang dimana berdasarkan data inputan 2800 sampel dan 7 kelas.

1. **Membuat LSTM Model**

   LSTM merupakan Long short term memory network (LSTM) yang dimana merupakan salah satu jenis dari Recurrent Neural Network (RNN) dimana dilakukan modifikasi dengan menambahkan memory cell yang dapat menyimpan informasi untuk jangka waktu yang lama [4]. LSTM diusulkan sebagai solusi untuk mengatasi terjadinya vanishing gradient pada RNN saat memproses data sequential yang panjang.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 6. Arsitektur LSTM [7]

Pada arsitektur LSTM ini diambil dari sampel hasil dari ekstrasi fitur MFCC yaitu 2800 sampel dan dimensi arraynya (40, 1 yang dimana akan ditaruh di layer pertama LSTM. Lalu layer selanjutnya yaitu Dense (64) dengan aktivasi relu untuk meminimalkan nilai kesalahan antara lapisan output dan kelas target lalu menambahkan dropout(0.2). Setelah itu layer dense dengan output 32 dan terakhir dengan output 7 dengan aktivasi softmax sebagai kategori pertama jika memilki klasifikasi binary.

3. Hasil dan Pembahasan

Pada penelitian kali ini saya menggunakan Recurrent Neural Networks (RNN) sebagai salah satu bentuk arsitektur Artificial Neural Networks (ANN) yang dirancang khusus untuk memproses data yang bersambung / berurutan (sequential data) [5]. Pengenalan suara atau speech recognition adalah salah satu aplikasi yang termasuk dalam arsitektur RNN.

Layer (type)	Output Shape	Param #
lstm_30 (LSTM)	(None, 256)	264192
dropout_62 (Dropout)	(None, 256)	0
dense_72 (Dense)	(None, 128)	32896
dropout_63 (Dropout)	(None, 128)	0
dense_73 (Dense)	(None, 64)	8256
dropout_64 (Dropout)	(None, 64)	0
dense_74 (Dense)	(None, 7)	455

Total params: 305,799
Trainable params: 305,799
Non-trainable params: 0

Gambar 7. Model sequential

- Dense = lapisan linier dimensi tunggal dengan unit tersembunyi
- Dropout = digunakan untuk menambahkan regularisasi ke data, menghindari over fitting & drop out sebagian kecil dari data
- Loss='sparse_categorical_crossentropy' = menghitung kerugian lintas-entropi antara label yang sebenarnya dan label yang diprediksi.
- Optimizer='adam' = secara otomatis menyesuaikan tingkat pembelajaran untuk model selama jumlah epochs

Klasifikasi merupakan termasuk tahapan penting karena pada tahapan adalah penentuan hasil dari proses ekstraksi fitur yang telah dilakukan sebelumnya. Klasifikasi emosi pada penelitian ini menggunakan metode Jaringan Syaraf Tiruan: Multilayer Perceptron (MLP).

3.1 Data Uji

Pada tahap selanjutnya yaitu menentukan parameter-parameter yang terbaik, untuk digunakan ke tahap kesimpulan hasil akhir. Penentuan parameter tersebut berada pada tabel dibawah ini :

Parameter	**Nilai**
Learning rate	0.1
Epochs	50
Layer	7
Batch Size	64
Validation_split	0.2
Validasi Akurasi terbaik	72.32

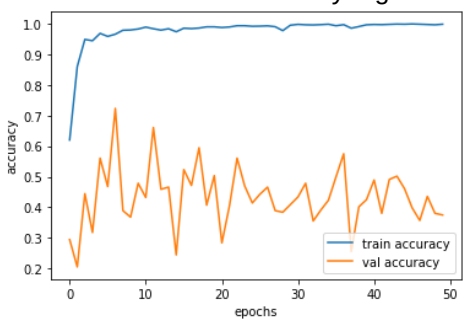
Tabel 2. Parameter Klasifikasi berdasarkan MLP

- batch_size=64 = jumlah data yang akan diproses per langkah
- epochs=50 = nomor iterasi untuk melatih model
- validasi_split=0.2 = melatih dan menguji persentase pemisahan
- Akurasi pelatihan dan akurasi validasi meningkatkan setiap iterasi
- akurasi validasi terbaik yang telah didapatkan yaitu 72.32 %

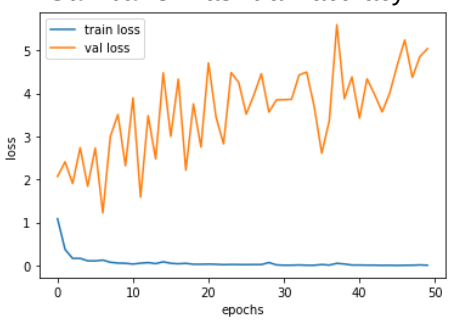
Jika diubah parameternya dalam menyesuaikan learning rate pada train model yang sudah dibuat dengan learning rate value nya yaitu 0.001 menjadikan hasil validasi akurasi hanya akan mendapatkan validasi akurasi dengan yang terbaik yaitu 57.14 % dan akan terus menurun seingiring bertambahnya Epochs karena learning rate menjadi terlalu besar.

Hasil melalui graphs plot

Pada tahap ini yaitu membuat plot hasil model yang dimana bagian yang dapat menyimpulkan hasil dari penelitian ini dengan menentukan validasi dari akurasi yang telah di uji sebelumnya.



Gambar 8. Hasil train acuracy



Gambar 9. Hasil train loss

Dalam penilitian ini didapatlan akurasi validasi akan menjadi sekitar 72,32% yang dimana lebih baik dalam konteks multiklasifikasi dasar. Pada penelitian ini dari perbedaan kalimat tidak dapat

p-ISSN: 2301-5373
e-ISSN: 2654-5101

mempengaruhi hasil akurasi melainkan beberapa faktor eksternal seperti noise dan emosi yang dibuat-buat oleh sample suara. Akurasi akhir dari model pada data latih pada model yang dibuat dengan menyesuaikan parameter learning rate dengan value yang beda mempunyai perbedaan yang signifikan karena konvergensi menjadi cukup lambat sampai model mengalami overfitting. Hal tersebut disebabkan oleh tidak seimbangnya kelas emosi pada dataset sehingga membuat model akan cenderung memprediksi kelas yang labelnya lebih banyak. Selain itu, kurangnya heterogenitas dari dataset yang membuat karakter kelas emosi lebih berbeda dari yang lainnya sehingga mampu mengurangi bias pada model agar tidak membuat model overfitting.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan pada kesempatan kali ini, Dari akurasi tersebut yang mencakup semua kelas emosi yang telah ditentukan sebelumnya dengan menggunakan metode MFCC, Bisa disimpulkan permasalahan overfitting masih terjadi, Maka dari itu memerlukannya strategi lain. Hal ini akan ditangani di masa mendatang.

Daftar Pustaka

[1] Fatan Kasyidi, Ridwan Ilyas, Nida Muthi Annisa, "Peningkatan Kemampuan Pengenalan Emosi melalui Suara dalam Bahasa Indonesia" *MIND (*Multimedia Artificial Intelligent Networking Database) Journal*,* vol 6, no 2, Hal 194, 2021.

[2] Raynaldy Arief, Nur Aviva Iriawan, Armin Lawi, “KLASIFIKASI AUDIO UCAPAN EMOSIONAL MENGGUNAKAN MODEL LSTM” KonferensiNasional Ilmu Komputer (KONIK), Hal 524, 2021.

[3] Barlian Henryranu Prasetio, Wijaya Kurniawan, Mochammad Hannats Hanafi Ichsan, " Pengenalan Emosi Berdasarkan Suara Menggunakan Algoritma HMM" (JTIIK) Jurnal Teknologi Informasi dan Ilmu Komputer. Vol 4, No 3, 2017.

[4] Yulistia Khoirotul Aini, Tri Budi Santoso, Titon Dutono, " Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia" (JKT) Jurnal Komputer Terapan, Vol 7, No 1, 2021.

[5] Angga Anjaini Sundawa, Aji Gautama Putrada, S.T., M.T., Novian Anggis Suwastika, S.T., M.T., “Implementasi dan Analisis Simulasi Deteksi Emosi Melalui Pengenalan Suara Menggunakan Mel-Frequency Cepstrum Coefficient dan Hidden Markov Model Berbasis IOT”, Universitas Telkom, 2019.

[6] Irham Sidik Permana , Youllia Indrawaty Nurhasanah , Andriana Zulkarnain, “IMPLEMENTASI METODE MFCC DAN DTW UNTUK PENGENALAN JENIS SUARA PRIA DAN WANITA’ MIND Journal, Vol 3, No 1, Hal 6, 2018.

[7] Xuan Hien Le, Hung Viet Ho, Giha Lee, Sungho Jung, “Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting” MDPI Journal, Department of Disaster Prevention and Environmental Engineering, Kyungpook National University, Korea.

This page is intentionally left blank.