JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Kombinasi Metode MFCC dan KNN dalam Pengenalan Emosi Manusia Melalui Ucapan

Putu Widya Eka Safitri[1], AAIN Eka Karyawati [a2]

[a] Program Studi Informatika, Universitas Udayana
Kuta Selatan, Badung, Bali, Indonesia
[1]widyaekasafitri79@gmail.com
[2]eka.karyawati@unud.ac.id

Abstract

Emotions are expressions that humans have in responding or responding to things that happen to themselves or in the environment around them, with these emotions humans are more able to express what they feel about a situation or event, these emotions can also be a means of communication other than language, because with emotions Of course, humans can know what happens to other humans around them. One of the human expressions to be able to communicate is by voice, sound can also be used to find out the type of emotion that is being experienced by the speaker. Mel-Frequency Cepstrum Coefficient (MFCC) is one of the feature extractions that is often used in the field of speech technology where in this feature extraction the human voice recording will be converted into a convolution matrix, namely a spectrogram or voice signal. K-Nearest Neighbor (K-NN) is a method that works by grouping new data based on the distance (neighborhood) from one data to the other. In the study of classical human emotions with speech using the K-Nearest Neighbor (K-NN) method, it is not appropriate to use this method because it only gets 50% accuracy.

Keywords: *Emotion, Voice, K-Nearest Neighbor, MFCC, Music Information Retrieval*

1. Pendahuluan

Emosi merupakan bagian dari manusia yang diekspresikan secara nyata (Gumelar et al., 2019). Emosi adalah ekspresi manusia dalam menunjukan reaksi terhadap suatu kejadian yang juga dapat digunakan sebagai alat komunikasi. Emosi banyak jenisnya, untuk jenis emosi yang biasa dilakukan seperti senang, takut, sedih, netral, marah dan jijik, sebenarnya banyak cara dalam mengekspresikan emosi seperti raut wajah, menulis, ucapan dan lain-lain.

Ucapan adalah cara manusia berkomunikasi dengan sesama manusia, dengan ucapan ini manusia lebih mudah dalam mengekspresikan emosinya (Helmiah et al., 2019). Ucapan merupakan sinyal kompleks yang berisi informasi. Manusia dapat berdialog untuk menyampaikan sesuatu dengan berbagai emosi yang dirasakan,

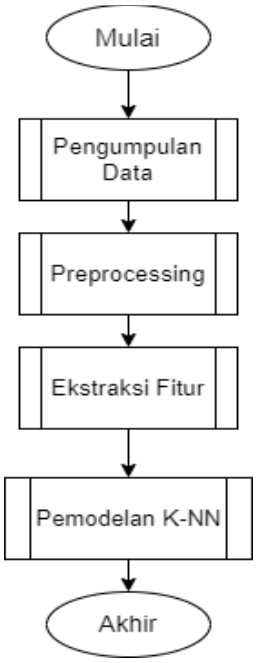
Pada penelitian ini akan dilakukan Klasifikasi emosi manusia dengan ucapan dengan metode *K-Nearest Neighbor (*K-NN) menggunakan ekstraksi fitur *Mel-Frequency Cepstrum Coefficient* (MFCC). *Mel-Frequency Cepstrum Coefficient (*MFCC) merupakan salah satu ekstraksi fitur yang sering digunakan dalam bidang *speech technologi* yang dimana pada ekstraksi fitur ini rekaman suara manusia akan diubah menjadi convolution matriks yaitu spectrogram atau sinyal suara dan nantinya akan diklasifikasi berdasarkan jenis emosinya.

K-Nearest Neighbor (KNN) dipilih karena bisa menyederhanakan perhitungan algoritma dan dapat mengefisiensikan waktu. *K-Nearest Neighbor* (KNN) adalah metode yang bekerja dengan cara mengelompokan data baru didasarkan pada jarak (bertetangga) dari satu data ke data yang

satunya (Fajar et al., 2019) Klasifikasi jenis emosi ini akan dilakukan dengan menggunakan algoritma *K-Nearest Neighbor* (KNN).

2. Metode Penelitian

Penelitian ini dilakukan untuk mengetahui jenis emosi manusia menggunakan *K-Nearest Neighbor* (K-NN), tahapan yang dilakukan pada penelitian ini yaitu pengumpulan data, preprocessing, ekstraksi fitur dan proses klasifikasi, dibawah ini merupakan alur dari penelitian



Gambar 1. Diagram Alir Pemodelan K-NN

Pada alur penelitian diatas, dimulai dengan pengumpulan data, yang dimana dataset pada tahap pengumpulan data tersebut bersumber dari Kaggle, lalu pada tahap Preprocessing akan dilakukan untuk membersihkan data, terdapat empat preprocessing yang diimplementasikan pada data adalah pengurangan noise, stretch data suara, shifting dan penyesuaian pitch, lalu pada ekstraksi fitur yang digunakan ekstraksi fitur *Mel-Frequency Cepstrum Coefficient (*MFCC) dan data suara akan diubah menjadi sinyal suara, dan yang terakhir pada tahap pemodelan *K-Nearest Neighbor* (K-NN) akan dicari akurasinya.

2.1 Tahap Pengumpulan Data

Dataset yang digunakan berjumlah 7.244 data yang dimana data yang diperankan oleh 91 aktor dengan pembagian 48 suara pria dan 43 suara wanita, yang rentang usianya 20 sampai 74 tahun dari berbagai etnis dan ras yang berbeda, didalam terdapat beberapa jenis emosi yaitu seperti marah, jijik, takut, senang, netral, dan sedih yang berbahasa Inggris.

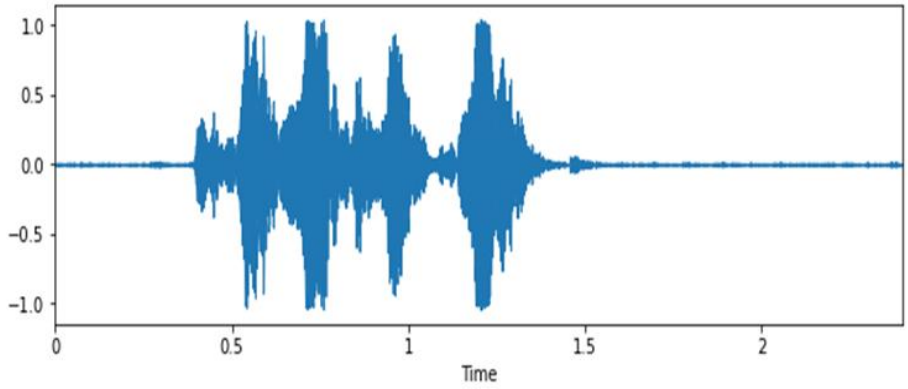
2.2 Tahap Preprocessing Ucapan

Pada proses ini, data yang dimiliki akan melalui tahap pembersihan data yang dimana akan dilakukan, untuk memaksimalkan data dan mengurangi kemungkinan error, dilakukan preprocessing pada data suara yang dibantu menggunakan library librosa dan numpy, beberapa *preprocessing* yang diimplementasikan pada data adalah pengurangan *noise*, *stretch* data suara,

shifting dan penyesuaian *pitch*. Sesuai penjelasan diatas, terdapat empat tahapan *noise reduction* untuk pengurangan noise, stretch untuk menghilangkan jeda pada data suara, shift yaitu menggeser waktu audio kami ke kiri atau ke kanan secara acak dengan persentase kecil, atau mengubah pitch atau kecepatan audio dengan jumlah kecil, pitch yaitu untuk menyesuaikan suara agar tidak terlalu rendah atau terlalu tinggi. Berikut merupakan hasil sinyal suara yang diuji coba pada penelitian ini setelah melalui tahap preprocessing:

a. Emosi Fear (Takut)

1.

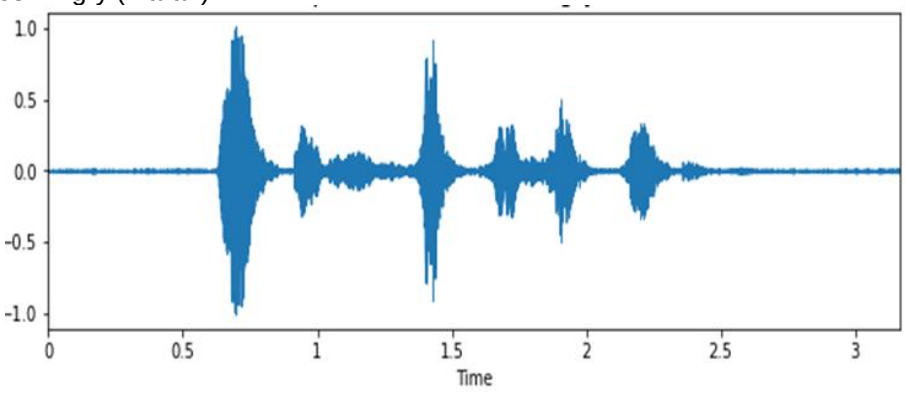


Gambar 2. Sinyal Suara Emosi Fear (Takut)

Emosi takut dirasakan apabila seseorang merasa terancam akan suatu hal yang sedang terjadi, hatinya akan merasa gelisah dan suara biasanya akan gemetar, suara meninggi tetapi memelan.

b. Emosi Angry (Marah)

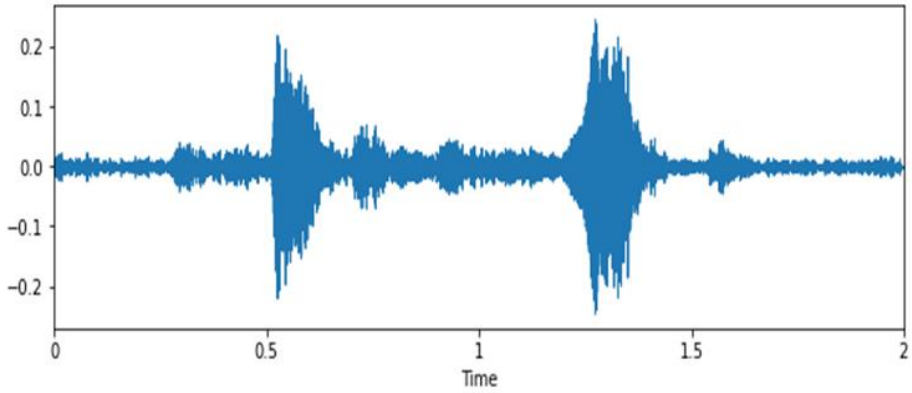
2.



Gambar 3. Sinyal Suara Emosi Angry (Marah)

Emosi marah ini akan terjadi ketika seseorang merasa harapan atau kehendaknya tidak terpenuhi dan tidak sesuai dengan harapannya

c. Emosi Happy (Senang)

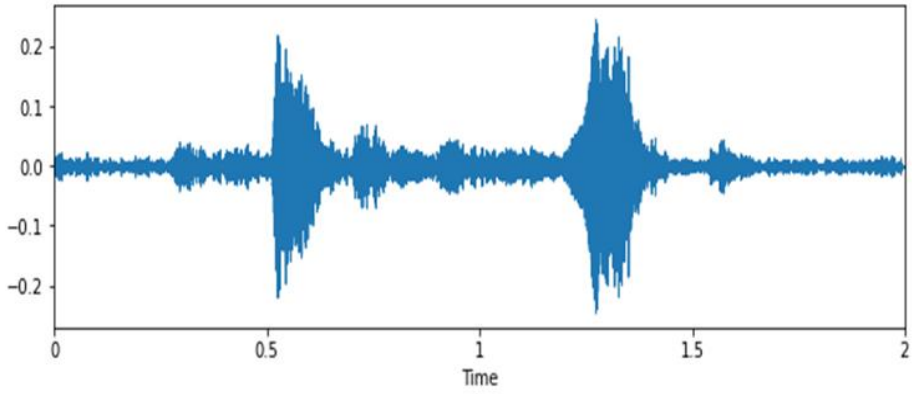


3.

Gambar 4. Sinyal Suara Emosi Happy (Senang)

Emosi senang ini muncul ketika ketika harapannya atau kehendaknya terpenuhi atau sesuatu yang menyenangkan hatinya terjadi

d. Emosi Sad (Sedih)

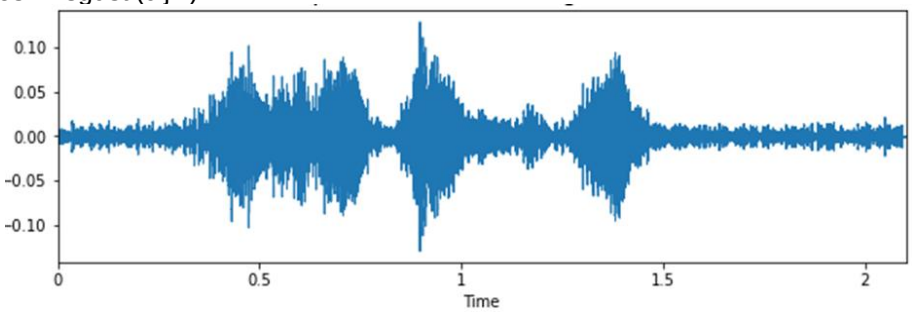


4.

Gambar 5. Sinyal Suara Emosi Sad (Sedih)

Emosi sedih ini terjadi ketika manusia merasa frustasi dan kecewa terhadap suatu hal atau seseorang, hatinya akan merasa kosong dan tidak puas akan sesuatu atau seseorang.

e. Emosi Disgust (Jijik)

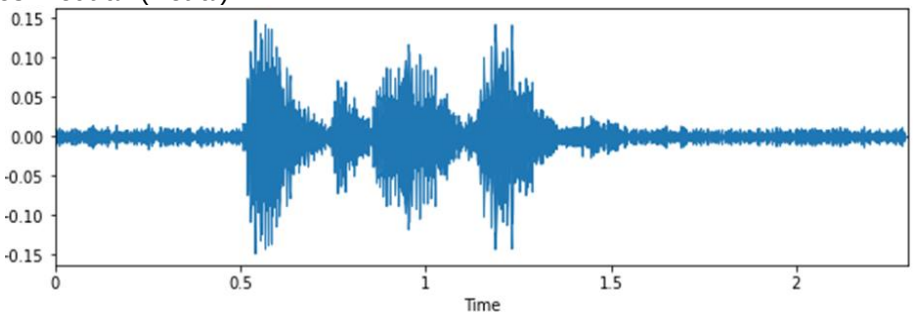


5.

Gambar 6. Sinyal Suara Emosi Disgust (Jijik)

Emosi ini terjadi apa bila manusia merasakan bahwa hal itu tidak seharusnya terjadi, dan timbul perasaan tidak ingin melihat, mendengar atau merasakan hal tersebut

f. Emosi Neutral (Netral)



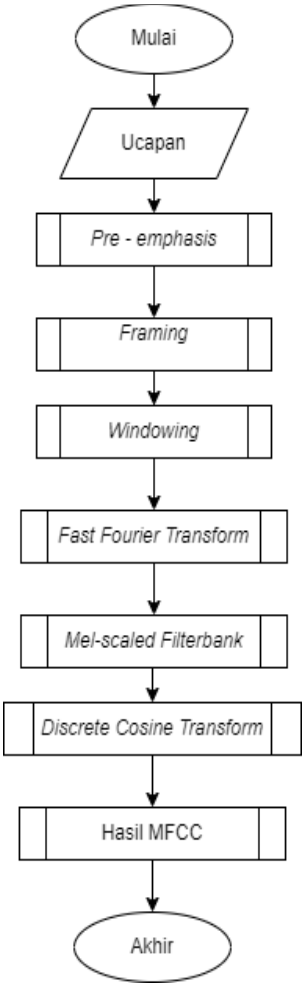
6.

Gambar 7. Sinyal Suara Emosi Neutral (Netral)

Emosi ini terjadi apabila tidak merasakan emosi-emosi diatas, tetapi hal ini juga dapat terjadi pada emosi baru yang tidak bisa dijelaskan.

2.3 Ekstraksi Fitur dengan MFCC

Pada penelitian dilakukan ekstraksi fitur *Mel-Frequency Cepstrum Coefficient (*MFCC) menggunakan library librosa, yang selanjutnya disimpan kedalam dataset. Pada ekstraksi fitur *Mel-Frequency Cepstrum Coefficient* (MFCC) menghasilkan 20 set data yang nantinya akan digunakan sebagai fiture untuk metode KNN.

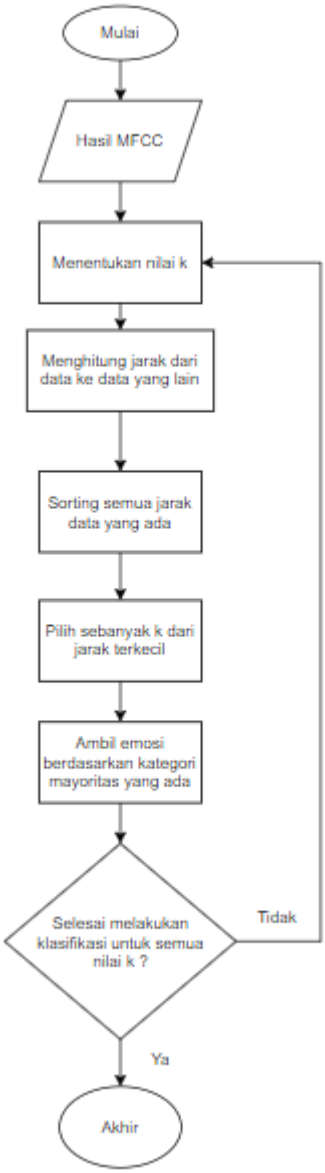


Gambar 8. Alur MFCC

Pada proses alur MFCC tersebut, dari inputan data yang telah melalui preprocessing yaitu data ucapan yang telah menjadi data numerik, lalu akan melalui tahapan Pre – emphasis Filtering yang merupakan salah satu jenis filter yang sering digunakan sebelum sebuah sinyal diproses lebih lanjut, lalu pada tahap framing yang bertujuan untuk memotong agar sinyal suara menjadi durasi yang lebih kecil dan stabil, lalu pada tahap windowing bertujuan mencegah terjadinya kebocoran spektral aliasing, pada tahap Fast Fourier Transform ini bertujuan untuk merubah sinyal dari domain waktu ke frekuensi agar sinyal dapat di proses dalam spectral substraksi, selanjutnya pada tahap Mel Scale Filterbank bertujuan untuk menyesuaikan sistem dengan pendengaran manusia, pada tahap Discrete Cosine Transform ini akan mengkompres sinyal yang akan dimasukkan ke frekuensi, dan setelah sinyal diubah ke frekuensi maka pada MFCC akan dilakukan penangkapan karakteristik dari suatu sinyal suara.

2.4 Pemodelan *K-Nearest Neighbor (*K-NN)

Pada penelitian Kombinasi Metode MFCC dan KNN dalam Pengenalan Emosi Manusia Melalui Ucapan, berikut ditampilkan pada gambar Flowchart alur penelitian ini.



Gambar 9. Flowchart Pemodelan K-NN

Pada Flowchart diatas alur dimulai dengan inputnya yaitu hasil dari MFCC, yang dimana akan dicari nilai k nya, lalu akan dihitung jarak dari data ke data yang sudah tersebar, setelah itu *sorting* semua jarak data, selanjutnya pilih data dengan jarak terkecil dan ambil emosi berdasarkan kategori mayoritas yang muncul. Setelah itu akan kembali ke tahap tanya apakah semua nilai k sudah diklasifikasi, jika sudah maka akan berakhir jika tidak akan diulangi proses tersebut sampai nilai k selesai di cari. Untuk mendapatkan tetangga terdekat didapatkan dari jarak terkecil antara data *input* dengan seluruh data *training* sebanyak nilai k

2.5 Tahap Evaluasi

Pada tahap evaluasi sistem, akan dicari akurasi dari penggunaan ekstraksi fitur MFCC dan metode KNN pada penelitian emosi manusia melalui ucapan.

Untuk perhitungan akurasi dilakukan menggunakan persamaan:

$$Akurasi = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

Keterangan :
TP = true positive
TN = true negative
FP = false positive
FN = false negative

3. Hasil dan Pembahasan

Tabel 1. Hasil Akurasi

k	**Akurasi**
3	50%
5	46,66%
7	45%
9	49,16%
11	49,16%

Tabel diatas merupakan jumlah k yang diuji, yaitu ditampilkan 6 model terbaik, akhirnya didapatkan bahwa dengan k yaitu 3 didapatkan nilai akurasi paling besar dari semua k yang diujikan yaitu 50%.

4. Kesimpulan

Berdasarkan penelitian yang telah ditulis dari pemodelan emosi melalui ucapan yang dibuat, dengan menggunakan ekstraksi fitur *Mel-Frequency Cepstrum Coefficient (*MFCC) yang dimana pada ekstraksi fitur *Mel-Frequency Cepstrum Coefficient* (MFCC) data numeric yang diambil yaitu berjumlah 20 set data dan dibuatkan grafik, dimasukan data baru yang tersebar acak dan data tersebut akan dikelompokkan dengan menggunakan metode *K-Nearest Neighbor* (K-NN) dengan

nilai k yaitu 3 sampai 101, yang dimana nilai k haruslah ganjil, setelah iterasi sebanyak 48, didapatkan bahwa dengan nilai k yaitu 3 dan 13 didapatkan akurasi terbaik yaitu 50%. Jadi dapat disimpulkan bahwa Pemodelan Emosi Manusia Melalui Ucapan Menggunakan Metode K-Nearest Neighbor (K-NN), kurang baik sehingga penelitian ini akan dilanjutkan di Tugas Akhir dengan metode yang berbeda untuk mendapatkan akurasi yang lebih tinggi.

Referensi

[1] Fajar Septria, Jangkung Raharjo, Nur Ibrahim, S.T.,M.T, " KLASIFIKASI EMOSI BERDASARKAN SINYAL SUARA MANUSIA MENGGUNAKAN METODE K-NEAREST NEIGHBOR (K-NN) " e-Proceeding of Engineering, Vol.6, No.2. Page 4130, 2019.

[2] Siti Helmiyah, Imam Riadi, Rusydi Umar, Abdullah Hanif, Anton Yudhana, Abdul Fadlil, "IDENTIFIKASI EMOSI MANUSIA BERDASARKAN UCAPAN MENGGUNAKAN METODE EKSTRAKSI CIRI LPC DAN METODE EUCLIDEAN DISTANCE " Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), Vol. 7, No. 6. Page 1177-1186, 2020.

[3] Yulistia Aini, Tri Budi Santoso, Titon Dutono. " Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia " Jurnal Politeknik Caltex Riau, Vol.7, No.1. Page 143 – 152, 2021.

[4] GUMELAR, A. B. et al. *"Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks", International Conference on Serious Games and Applications for Health (SeGAH). IEEE, Page 1– 8. 2019*

[5] Anjani Reddy J, Dr. Shiva G. "*Emotion Recognition from Speech Using MLP AND KNN* ", RESEARCH ARTICLE. Vol. 11, (Series-II) Page 34-38. 2021

[6] Steven R. Livingstone, Frank A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", PLoS ONE 13(5): e0196391. 2018.