JITSI Vol 1 No. 1 (2026)
Journal of Information Technology and System Integration
https://journal2.ummat.ac.id/index.php/jitsi

Deteksi Emosi dari Suara Menggunakan MFCC dan Algoritma SVM

Riota Adi Saputra[*],Alfian Haris[2],Bayu Sugarda [3] ,Pandi Agustia[4]
[1,2]Sistem dan Teknologi Informasi, Universitas Muhammadiyah Mataram
[3,4]Sistem dan Teknologi Informasi, Universitas Muhammadiyah Mataram
email: xxx@xxxxx.com[1*]

Abstrak: Speech Emotion Recognition (SER) is a field of artificial intelligence that aims to identify a speaker's emotional state based on speech characteristics. The ability to recognize emotions from speech plays an important role in developing adaptive human-computer interaction systems, including virtual assistants, automated customer services, and digital healthcare applications. This study proposes a speech emotion detection system using Mel-Frequency Cepstral Coefficients (MFCC) feature extraction and the Support Vector Machine (SVM) classification algorithm. The dataset employed is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), consisting of 1,440 audio recordings representing eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The research process includes audio preprocessing through pre-emphasis filtering and silence removal, followed by the extraction of MFCC, Delta MFCC, Delta-Delta MFCC, and complementary acoustic features, namely Zero Crossing Rate (ZCR), Root Mean Square (RMS) Energy, and Spectral Centroid. Feature values were summarized using mean and standard deviation statistics, resulting in 84 features for each audio sample. Feature normalization was performed using StandardScaler before classification with an SVM model employing a Radial Basis Function (RBF) kernel. Experimental results showed that the proposed model achieved a training accuracy of 97.31% and a testing accuracy of 64.58%. Furthermore, 5-Fold Cross Validation produced an average accuracy of 67.01% with a standard deviation of 1.89%, indicating relatively stable classification performance. Confusion matrix analysis revealed that the emotions angry, calm, and disgust were recognized more accurately than other emotion categories. The findings demonstrate that the combination of MFCC and SVM is effective for speech emotion classification, although signs of overfitting remain present, particularly among emotion classes with similar acoustic characteristics.

Keywords: Speech Emotion Recognition, MFCC, Support Vector Machine, RAVDESS, Speech Emotion Detection.

Abstrak: Speech Emotion Recognition (SER) merupakan salah satu bidang penelitian dalam kecerdasan buatan yang bertujuan mengenali kondisi emosional seseorang berdasarkan karakteristik suara. Kemampuan mengenali emosi dari suara memiliki peran penting dalam pengembangan sistem interaksi manusia dan komputer yang lebih adaptif, seperti asisten virtual, layanan pelanggan otomatis, serta aplikasi kesehatan digital. Penelitian ini bertujuan membangun sistem deteksi emosi suara menggunakan metode ekstraksi fitur Mel-Frequency Cepstral Coefficients (MFCC) dan algoritma klasifikasi Support Vector Machine (SVM). Dataset yang digunakan adalah Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) yang terdiri dari 1.440 file audio dengan delapan kategori emosi, yaitu neutral, calm, happy, sad, angry, fearful, disgust, dan surprised. Tahapan penelitian meliputi preprocessing berupa pre-emphasis dan penghilangan segmen senyap, ekstraksi fitur MFCC, Delta MFCC, Delta-Delta MFCC, serta fitur komplementer berupa Zero Crossing Rate (ZCR), Root Mean Square (RMS) Energy, dan Spectral Centroid. Seluruh fitur diringkas menggunakan nilai rata-rata dan standar deviasi sehingga menghasilkan 84 fitur untuk setiap sampel audio. Selanjutnya dilakukan normalisasi menggunakan StandardScaler dan klasifikasi menggunakan SVM dengan kernel Radial Basis Function (RBF). Hasil penelitian menunjukkan bahwa model memperoleh akurasi pelatihan sebesar 97,31% dan akurasi pengujian sebesar 64,58%. Evaluasi menggunakan 5-Fold Cross Validation menghasilkan rata-rata akurasi sebesar 67,01% dengan standar deviasi 1,89%, yang menunjukkan performa model cukup stabil. Analisis confusion matrix menunjukkan bahwa emosi angry, calm, dan disgust memiliki tingkat pengenalan yang lebih baik dibandingkan kategori emosi lainnya. Hasil penelitian membuktikan bahwa kombinasi MFCC dan SVM mampu melakukan klasifikasi emosi suara dengan performa yang cukup baik, meskipun masih ditemukan indikasi overfitting pada beberapa kategori emosi yang memiliki karakteristik akustik serupa.

Kata Kunci: Speech Emotion Recognition, MFCC, SVM, RAVDESS, Deteksi Emosi Suara.

PENDAHULUAN

Perkembangan teknologi kecerdasan buatan telah mendorong peningkatan penelitian dalam bidang pengenalan suara, khususnya pada Speech Emotion Recognition (SER) yang bertujuan untuk mengidentifikasi kondisi emosional seseorang melalui sinyal ucapan. Emosi yang terkandung dalam suara manusia dapat memberikan informasi penting mengenai keadaan psikologis, niat komunikasi, maupun respons afektif pembicara. Untuk dapat mengenali emosi secara otomatis,

sistem SER memerlukan proses ekstraksi fitur yang mampu merepresentasikan karakteristik suara secara efektif serta metode klasifikasi yang mampu membedakan berbagai kategori emosi. Salah satu teknik ekstraksi fitur yang banyak digunakan adalah Mel Frequency Cepstral Coefficients (MFCC) karena mampu meniru mekanisme persepsi pendengaran manusia dalam menangkap informasi spektral suara. Sementara itu, Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin yang dikenal memiliki kemampuan klasifikasi yang baik pada data berdimensi tinggi dan jumlah sampel yang relatif terbatas, sehingga banyak diterapkan pada berbagai penelitian pengenalan suara dan emosi. Penelitian mengenai pengenalan emosi dari suara telah berkembang pesat dalam beberapa tahun terakhir. Trigeorgis dkk. menunjukkan bahwa penggunaan pendekatan deep learning berbasis spektrogram mampu meningkatkan kemampuan sistem dalam mengenali pola emosi yang kompleks secara otomatis tanpa memerlukan rekayasa fitur yang berlebihan. Hasil penelitian tersebut menunjukkan bahwa representasi spektral suara mengandung informasi penting yang berkaitan dengan ekspresi emosional pembicara sehingga dapat dimanfaatkan untuk meningkatkan performa sistem Speech Emotion Recognition. Temuan ini memperlihatkan bahwa karakteristik akustik suara memiliki hubungan yang erat dengan kondisi emosi seseorang dan menjadi dasar penting dalam pengembangan sistem pengenalan emosi modern (Cortes and Vapnik 1995).

Selain pemanfaatan deep learning, penelitian lain juga menunjukkan pentingnya proses pembelajaran representasi fitur yang efektif dalam meningkatkan akurasi klasifikasi emosi. Satt dkk. mengemukakan bahwa metode unsupervised representation learning mampu menghasilkan representasi fitur yang lebih informatif dibandingkan pendekatan konvensional sehingga dapat meningkatkan performa sistem Speech Emotion Recognition. Di sisi lain, perkembangan terbaru yang diperkenalkan melalui model emotion2vec menunjukkan bahwa pendekatan self-supervised learning mampu mempelajari representasi emosi secara lebih umum dan robust dari data suara dalam jumlah besar. Hasil penelitian tersebut membuktikan bahwa kualitas representasi fitur merupakan faktor penting yang menentukan keberhasilan proses klasifikasi emosi pada sistem pengenalan suara (Satt, Rozenberg, and Hoory 2017).

Pada aspek ekstraksi fitur dan klasifikasi, MFCC masih menjadi salah satu metode yang paling banyak digunakan dalam pengolahan sinyal suara karena kemampuannya dalam merepresentasikan informasi frekuensi yang relevan dengan persepsi pendengaran manusia. Penelitian mengenai MFCC menunjukkan bahwa metode ini mampu menghasilkan fitur yang efektif untuk berbagai aplikasi pemrosesan ucapan, termasuk pengenalan emosi. Sementara itu, algoritma Support Vector Machine yang diperkenalkan oleh Cortes dan Vapnik terbukti memiliki kemampuan klasifikasi yang baik melalui pencarian hyperplane optimal yang memaksimalkan margin antar kelas. Kombinasi MFCC sebagai metode ekstraksi fitur dan SVM sebagai metode klasifikasi telah banyak digunakan karena menawarkan keseimbangan antara kompleksitas komputasi dan tingkat akurasi yang memadai, terutama pada penelitian dengan sumber daya komputasi yang terbatas (Neumann and Vu 2019).

Meskipun berbagai penelitian terkini telah menunjukkan keberhasilan penggunaan deep learning dan self-supervised learning dalam pengenalan emosi suara, sebagian besar metode tersebut memerlukan jumlah data yang besar serta sumber daya komputasi yang tinggi. Di sisi lain, masih diperlukan kajian mengenai efektivitas pendekatan yang lebih sederhana namun tetap mampu menghasilkan performa yang baik pada proses klasifikasi emosi suara. Selain itu, penelitian yang secara khusus mengintegrasikan MFCC sebagai metode ekstraksi fitur dan SVM sebagai algoritma klasifikasi pada skenario deteksi emosi masih relevan untuk dieksplorasi sebagai alternatif yang lebih efisien. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem deteksi emosi dari suara menggunakan ekstraksi fitur MFCC dan algoritma Support Vector Machine (SVM), serta mengevaluasi kemampuan metode tersebut dalam mengklasifikasikan berbagai kategori emosi secara akurat sebagai pendekatan cerdas dalam pengenalan suara (Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen 2023).

Perkembangan teknologi kecerdasan buatan dan Human-Computer Interaction (HCI) telah mendorong peningkatan penggunaan sistem berbasis suara dalam berbagai bidang kehidupan. Berbagai teknologi seperti asisten virtual, layanan pelanggan otomatis, sistem chatbot suara, perangkat pintar, serta aplikasi kesehatan digital semakin mengandalkan kemampuan komputer untuk memahami dan merespons ucapan manusia secara alami. Dalam implementasinya, pemahaman terhadap isi ucapan saja sering kali belum cukup untuk menghasilkan interaksi yang efektif. Sistem juga perlu memahami aspek emosional yang terkandung dalam suara agar dapat memberikan respons yang lebih sesuai dengan kondisi pengguna. Oleh karena itu, kemampuan mendeteksi emosi dari suara menjadi salah satu komponen penting dalam pengembangan sistem interaksi manusia dan komputer yang lebih cerdas dan adaptif (de Lope and Graña 2023).

Informasi emosi yang terkandung dalam suara dapat dimanfaatkan dalam berbagai aplikasi, seperti pemantauan kesehatan mental, sistem pembelajaran adaptif, layanan pelanggan otomatis, analisis perilaku pengguna, hingga sistem keamanan. Karakteristik suara seperti nada, intensitas, kecepatan bicara, dan pola frekuensi diketahui mengalami perubahan sesuai dengan kondisi emosional pembicara. Namun, proses identifikasi emosi secara manual memerlukan waktu dan sangat bergantung pada subjektivitas manusia. Oleh karena itu, diperlukan suatu sistem otomatis yang mampu mengenali emosi secara objektif dan konsisten melalui analisis sinyal suara. Pengembangan sistem tersebut menjadi salah satu fokus utama dalam bidang Speech Emotion Recognition (SER) yang saat ini terus berkembang seiring meningkatnya kebutuhan terhadap teknologi interaksi yang lebih natural dan berpusat pada pengguna (Haoxing and System n.d.).

JITSI Vol 1 No. 1 (2026)
Journal of Information Technology and System Integration
https://journal2.ummat.ac.id/index.php/jitsi

TINJAUAN PUSTAKA
METODE

A. Alur Penelitian

Penelitian ini menggunakan pendekatan klasifikasi sinyal suara untuk mengenali emosi pembicara secara otomatis. Alur penelitian terdiri dari enam tahap utama, yaitu: akusisi data, eksplorasi data, preprocessing, ekstraksi fitur MFCC, normalisasi, dan klasifikasi menggunakan SVM. Gambaran umum alur penelitian ditunjukkan pada Gambar 1.

Gambar 1. Alur penelitian.

B. Akusisi Data

Penelitian ini menggunakan dataset RAVDESS (Ryerson Audio-Visual Database of Emotional Speech), yaitu basis data audio-visual emosional yang bersifat terbuka dan tersedia secara publik, termasuk melalui platform Kaggle. Dataset ini dikembangkan oleh Livingstone dan Russo, memuat 24 aktor profesional (12 wanita dan 12 pria) yang melafalkan pernyataan dengan aksen Amerika Utara yang netral. Dalam penelitian ini, digunakan subset berupa 1.440 file rekaman audio berformat .wav yang mencakup delapan kategori emosi, yaitu: neutral, calm, happy, sad, angry, fearful, disgust, dan surprised. Setiap ekspresi diproduksi dalam dua tingkat intensitas emosional, dengan tambahan ekspresi netral. Penamaan setiap file mengikuti format kode terstruktur yang memuat informasi modalitas, saluran vokal, emosi, intensitas, pernyataan, repetisi, dan nomor aktor, sehingga proses pelabelan dapat dilakukan secara otomatis tanpa memerlukan anotasi manual.

C. Preprocessing

Tahap pertama adalah pre-emphasis, yakni penerapan filter high-pass orde pertama untuk memperkuat komponen frekuensi tinggi dalam sinyal. Filter ini bekerja dengan mengurangkan sebagian nilai sampel sebelumnya dari sampel saat ini, sehingga secara efektif memperkuat perbedaan antar sampel yang lebih menonjol pada frekuensi tinggi. Penerapan filter ini bermanfaat untuk menyeimbangkan spektrum frekuensi, meningkatkan Signal-to-Noise Ratio (SNR), serta menghindari masalah numerik pada saat komputasi FFT, sehingga fitur berbasis spektrum seperti MFCC maupun mel-spectrogram dapat diekstraksi secara lebih optimal. Nilai koefisien alpha yang digunakan adalah 0,97, dengan nilai tipikal yang umum digunakan dalam speech processing berkisar antara 0,95 hingga 0,97 (Fayek 2016). Filter ini bekerja berdasarkan persamaan:

$$y(t) = x(t) - \alpha x(t-1) \quad (1)$$

Tahap kedua adalah penghapusan segmen senyap (silence trimming). Penghapusan jeda senyap pada sinyal ucapan terbukti memberikan hasil Speech Emotion Recognition (SER) yang lebih baik pada model deep learning, karena segmen senyap tidak mengandung informasi ucapan yang relevan untuk pelatihan model. Proses ini bekerja dengan membandingkan energi tiap segmen sinyal terhadap puncak amplitudo keseluruhan — segmen yang energinya berada di bawah nilai threshold akan dipotong (Abubakar 2019). Nilai threshold yang digunakan dalam penelitian ini adalah 30 dB. Langkah ini bertujuan mengurangi durasi yang tidak relevan, menghilangkan background noise di awal dan akhir rekaman, serta mencegah model belajar dari segmen kosong yang berpotensi menurunkan performa klasifikasi.

D. Ektraksi Fitur MFCC

Mel-Frequency Cepstral Coefficients (MFCC) merupakan metode ekstraksi fitur yang banyak digunakan dalam pengolahan sinyal suara. Metode ini dirancang untuk merepresentasikan karakteristik spektral sinyal berdasarkan cara kerja sistem pendengaran manusia. MFCC diperoleh melalui serangkaian proses yang meliputi transformasi sinyal ke domain frekuensi, pemetaan frekuensi ke skala Mel, perhitungan logaritma energi, dan transformasi kosinus diskrit (DCT) (Sasongko et al. 2023). Hasil dari proses tersebut berupa sekumpulan koefisien yang dapat menggambarkan karakteristik suara secara ringkas. Hubungan antara frekuensi linear dan skala Mel dinyatakan sebagai berikut:

$$\left[Mel(f) = 2595\left(1 + \frac{f}{700}\right)\right] \quad (2)$$

Koefisien MFCC diperoleh menggunakan Discrete Cosine Transform (DCT) yang dirumuskan sebagai berikut:

$$\left[C(n) = \sum_{m=1}^{M} \quad S(m)\ cos\ cos\ \left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right]\right] \qquad (3)$$

1. Delta MFCC

Delta MFCC merupakan turunan pertama dari koefisien MFCC yang digunakan untuk menggambarkan perubahan karakteristik spektral suara dari waktu ke waktu. Fitur ini memberikan informasi temporal yang tidak dapat direpresentasikan oleh MFCC statis sehingga mampu menggambarkan dinamika ucapan secara lebih baik. Oleh karena itu, Delta MFCC sering digunakan untuk meningkatkan kemampuan sistem dalam membedakan pola suara yang memiliki karakteristik serupa. Perhitungan Delta MFCC dapat dinyatakan sebagai berikut.

$$\left[\Delta c_t = \frac{\sum_{n=1}^{N} \quad n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} \quad n^2}\right] \quad (4)$$

2. Delta-Delta MFCC

Delta-Delta MFCC atau acceleration coefficient merupakan turunan kedua dari koefisien MFCC. Fitur ini digunakan untuk menggambarkan percepatan perubahan karakteristik spektral suara sehingga dapat memberikan informasi temporal yang lebih rinci dibandingkan Delta MFCC. Penggunaan Delta-Delta MFCC memungkinkan sistem mengenali pola perubahan suara yang berkaitan dengan variasi emosi dalam ucapan. Perhitungan Delta-Delta MFCC dirumuskan sebagai berikut.

$$\left[\Delta^2 c_t = \frac{\sum_{n=1}^{N} \quad n(\Delta c_{t+n} - \Delta c_{t-n})}{2\sum_{n=1}^{N} \quad n^2}\right] \quad (5)$$

3. Fitur Komplementer

Selain MFCC, penelitian ini juga menggunakan beberapa fitur akustik tambahan untuk melengkapi representasi karakteristik suara. Fitur yang digunakan meliputi Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), dan Spectral Centroid. Ketiga fitur tersebut dipilih karena mampu memberikan informasi mengenai perubahan sinyal, energi suara, dan distribusi frekuensi yang tidak sepenuhnya direpresentasikan oleh MFCC.

a. Zero Crossing Rate

Zero Crossing Rate merupakan ukuran jumlah perubahan tanda sinyal dari positif ke negatif atau sebaliknya dalam suatu interval waktu. Nilai ZCR sering digunakan untuk menggambarkan tingkat perubahan sinyal dan karakteristik frekuensi suatu suara.

$$\left[ZCR = \frac{1}{2N}\sum_{n=1}^{N} \quad |sign(x_n) - sign(x_{n-1})|\right] \quad (6)$$

b. Root Mean Square Energy

Root Mean Square Energy digunakan untuk mengukur energi rata-rata suatu sinyal suara. Fitur ini dapat menggambarkan tingkat kekuatan atau intensitas suara yang dihasilkan oleh pembicara.

$$\left[RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \quad x_n^2}\right] \quad (7)$$

c. Spectral Centroid

Spectral Centroid merupakan fitur yang menunjukkan pusat massa distribusi spektrum frekuensi suatu sinyal. Nilai Spectral Centroid sering digunakan untuk menggambarkan tingkat kecerahan (brightness) suara.

$$SC = \frac{\sum_{k=1}^{K} \quad f_k X_k}{\sum_{k=1}^{K} \quad X_k} \quad (8)$$

d. Pembentukan Vektor Fitur

Setelah seluruh fitur diekstraksi, nilai-nilai fitur diringkas menggunakan ukuran statistik berupa mean dan standard deviation untuk menghasilkan representasi berdimensi tetap pada setiap file audio. Vektor fitur yang terbentuk selanjutnya digunakan sebagai masukan pada algoritma Support Vector Machine (SVM) untuk proses klasifikasi emosi suara.

E. Normalisasi

Normalisasi fitur dilakukan menggunakan StandardScaler (z-score normalization), dengan transformasi:

$$z = \frac{x-\mu}{\sigma} \quad (9)$$

di mana μ adalah rata-rata dan σ adalah standar deviasi fitur pada data latih. Proses fitting scaler dilakukan hanya pada data latih, kemudian diterapkan (transform) pada data latih maupun data uji untuk mencegah data leakage (George and Muhamed Ilyas 2026). Normalisasi ini diperlukan agar seluruh fitur berada pada skala yang sama, sehingga performa model SVM tidak didominasi oleh fitur dengan rentang nilai yang besar.

F. Support Vector Machine (SVM)

Setelah proses ekstraksi fitur selesai dilakukan, tahap berikutnya adalah klasifikasi emosi suara menggunakan algoritma Support Vector Machine (SVM). SVM merupakan algoritma pembelajaran terawasi (supervised learning) yang bekerja dengan mencari hyperplane optimal untuk memisahkan data dari kelas yang berbeda. Hyperplane tersebut dipilih sedemikian rupa sehingga menghasilkan margin maksimum terhadap titik data terdekat yang disebut support vector (Cortes and Vapnik 1995). Pada penelitian ini, SVM digunakan untuk mengklasifikasikan fitur suara ke dalam beberapa kategori emosi. Karena permasalahan yang dihadapi merupakan klasifikasi multikelas, digunakan pendekatan One-vs-Rest (OvR) yang membangun satu model klasifikasi untuk setiap kelas terhadap seluruh kelas lainnya (Rifkin, R., & Klautau 2004). Tujuan utama SVM adalah mencari hyperplane optimal yang dapat memaksimalkan margin antar kelas. Hyperplane tersebut dapat dinyatakan sebagai:

$$[w^T x + b = 0] \quad (10)$$

Proses optimasi SVM bertujuan meminimalkan fungsi berikut:

$$\frac{1}{2} w^T w \quad (11)$$

dengan syarat:

$$[y_i(w^T x_i + b) \geq 1] \quad (12)$$

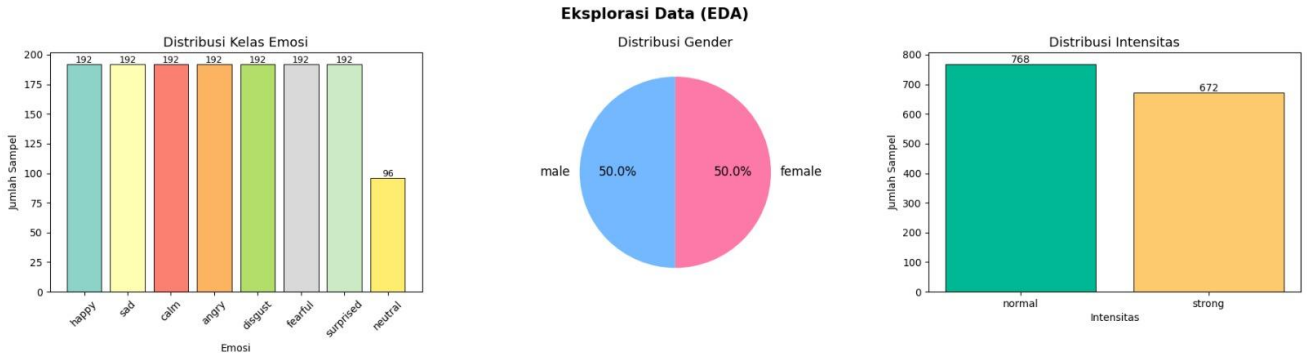
Karena data emosi suara memiliki pola yang kompleks dan tidak selalu dapat dipisahkan secara linear, penelitian ini menggunakan kernel Radial Basis Function (RBF). Kernel RBF berfungsi memetakan data ke ruang berdimensi lebih tinggi sehingga batas antar kelas dapat dipisahkan dengan lebih baik (Milton, Sharmy Roy, and Tamil Selvi 2013). Persamaan kernel RBF dinyatakan sebagai:

$$K(x_i, x_j) = exp\ exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (13)$$

HASIL DAN PEMBAHASAN

dataset RAVDESS memiliki distribusi data yang cukup baik untuk digunakan dalam penelitian klasifikasi emosi suara. Sebagian besar kelas emosi memiliki jumlah sampel yang sama sehingga model tidak cenderung mempelajari satu kelas tertentu secara berlebihan. Selain itu, distribusi gender yang seimbang antara laki-laki dan perempuan membantu meningkatkan kemampuan generalisasi model terhadap berbagai karakteristik suara. Meskipun terdapat perbedaan jumlah sampel pada kelas neutral dan distribusi intensitas suara, kondisi tersebut masih berada dalam batas yang dapat diterima dan tidak memberikan ketimpangan yang signifikan terhadap proses pelatihan model klasifikasi. Hasil analisis ini menunjukkan bahwa dataset layak digunakan untuk tahap ekstraksi fitur MFCC dan pelatihan model SVM pada penelitian deteksi emosi dari suara.

JITSI Vol 1 No. 1 (2026)
Journal of Information Technology and System Integration
https://journal2.ummat.ac.id/index.php/jitsi



Pada tahap preprocessing dilakukan pre-emphasis dan penghilangan bagian hening (silence removal). Pre-emphasis digunakan untuk memperkuat komponen frekuensi tinggi yang umumnya memiliki energi lebih rendah dibanding frekuensi rendah. Setelah itu dilakukan penghilangan bagian hening menggunakan metode trimming sehingga hanya bagian sinyal yang mengandung informasi suara yang diproses lebih lanjut.

Tahap ekstraksi fitur menggunakan MFCC sebanyak 13 koefisien yang dikombinasikan dengan Delta MFCC dan Delta-Delta MFCC. Selain itu ditambahkan fitur pendukung berupa Zero Crossing Rate, RMS Energy, dan Spectral Centroid untuk menangkap karakteristik temporal dan spektral suara. Hasil ekstraksi menghasilkan vektor fitur sebanyak 84 fitur untuk setiap sampel audio.

Fitur	Mean	Standar Deviasi	Total Fitur
MFCC (13)	13	13	26
Delta MFCC (13)	13	13	26
Delta-Delta MFCC (13)	13	13	26
Zero Crossing Rate	1	1	2
RMS Energy	1	1	2
Spectral Centroid	1	1	2
Total		84	

Setelah proses ekstraksi fitur selesai dilakukan, tahap berikutnya adalah label encoding. Tahap ini bertujuan untuk mengubah label emosi yang masih berbentuk teks (string) menjadi representasi numerik agar dapat diproses oleh algoritma klasifikasi Support Vector Machine (SVM). Proses konversi dilakukan menggunakan metode Label Encoder dari pustaka Scikit-learn. Setiap kategori emosi diberikan nilai numerik yang unik sehingga model dapat mengenali dan membedakan setiap kelas emosi selama proses pelatihan maupun pengujian. Hasil konversi label emosi ditunjukkan pada Tabel 4.X.

Label Kelas Emosi pada Dataset

Emosi	Label
Neutral	0
Calm	1
Happy	2
Sad	3
Angry	4

Fearful	5
Disgust	6
Surprised	7

dataset RAVDESS memiliki distribusi data yang cukup baik untuk digunakan dalam penelitian klasifikasi emosi suara. Sebagian besar kelas emosi memiliki jumlah sampel yang sama sehingga model tidak cenderung mempelajari satu kelas tertentu secara berlebihan. Selain itu, distribusi gender yang seimbang antara laki-laki dan perempuan membantu meningkatkan kemampuan generalisasi model terhadap berbagai karakteristik suara. Meskipun terdapat perbedaan jumlah sampel pada kelas neutral dan distribusi intensitas suara, kondisi tersebut masih berada dalam batas yang dapat diterima dan tidak memberikan ketimpangan yang signifikan terhadap proses pelatihan model klasifikasi. Hasil analisis ini menunjukkan bahwa dataset layak digunakan untuk tahap ekstraksi fitur MFCC dan pelatihan model SVM pada penelitian deteksi emosi dari suara.

Setelah data fitur hasil ekstraksi MFCC dinormalisasi menggunakan StandardScaler, tahap selanjutnya adalah melakukan pelatihan model klasifikasi menggunakan algoritma Support Vector Machine (SVM). Pada penelitian ini digunakan model SVM Baseline dengan parameter bawaan (default) sebagai acuan awal untuk mengukur performa klasifikasi. Model SVM menggunakan kernel Radial Basis Function (RBF) dengan nilai parameter C = 1,0 dan gamma = scale. Selain itu, digunakan pendekatan One-Versus-Rest (OvR) untuk menangani klasifikasi multikelas yang terdiri dari delapan kategori emosi.

Proses pelatihan dilakukan menggunakan data latih (training data) yang telah melalui proses normalisasi. Setelah model selesai dilatih, dilakukan pengujian menggunakan data uji (testing data) untuk mengetahui kemampuan model dalam mengenali emosi pada data yang belum pernah dilihat sebelumnya. Evaluasi kinerja model dilakukan dengan menghitung nilai akurasi pada data pelatihan dan data pengujian.

Kelas Emosi	Precision	Recall	F1-Score	Support
Angry	0.8095	0.8947	0.8500	38
Calm	0.7838	0.7632	0.7733	38
Disgust	0.7632	0.7632	0.7632	38
Fearful	0.7647	0.6667	0.7123	39
Happy	0.7353	0.6410	0.6849	39
Neutral	0.4706	0.4211	0.4444	19
Sad	0.5652	0.6842	0.6190	38
Surprised	0.7500	0.769	0.7595	39

KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini berhasil mengembangkan sistem Speech Emotion Recognition (SER) menggunakan kombinasi metode ekstraksi fitur Mel-Frequency Cepstral Coefficients (MFCC) dan algoritma klasifikasi Support Vector Machine (SVM) untuk mengenali delapan kategori emosi pada dataset RAVDESS. Proses ekstraksi fitur yang menggabungkan MFCC, Delta MFCC, Delta-Delta MFCC, serta fitur komplementer berupa Zero Crossing Rate (ZCR), RMS Energy, dan Spectral Centroid menghasilkan representasi suara yang mampu menggambarkan karakteristik spektral dan temporal sinyal ucapan secara efektif.

Hasil pengujian menunjukkan bahwa model SVM dengan kernel RBF memperoleh akurasi pelatihan sebesar 97,31% dan akurasi pengujian sebesar 64,58%. Evaluasi menggunakan metode 5-Fold Cross Validation menghasilkan rata-rata akurasi sebesar 67,01% dengan standar deviasi 1,89%, yang menunjukkan bahwa model memiliki tingkat stabilitas yang cukup baik dalam melakukan klasifikasi emosi suara. Berdasarkan analisis performa per kelas, emosi angry, calm, dan disgust memiliki tingkat pengenalan yang lebih tinggi dibandingkan kategori emosi lainnya.

Secara keseluruhan, penelitian ini membuktikan bahwa kombinasi MFCC dan SVM dapat digunakan sebagai pendekatan yang efektif dan efisien untuk deteksi emosi dari suara. Namun, perbedaan yang cukup besar antara akurasi pelatihan dan pengujian mengindikasikan adanya kecenderungan overfitting, sehingga diperlukan pengembangan lebih lanjut melalui optimasi hyperparameter, seleksi fitur, reduksi dimensi, maupun penerapan metode deep learning untuk meningkatkan kemampuan generalisasi model pada data yang lebih beragam.

DAFTAR PUSTAKA

Abubakar, Abdulrahman et al. 2019. “Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition.” https://arxiv.org/abs/1907.06078.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20(3): 273–97.

Fayek, Haytham M. 2016. “Speech Processing for Machine Learning: Filter Banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What’s In-Between.” https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

George, Swapna Mol, and P Muhamed Ilyas. 2026. “Assessing the Effectiveness of Feature Normalization and Dataset Quality in Speech Emotion Recognition Across Diverse Emotional and Linguistic Contexts.” *Circuits, Systems, and Signal Processing* 45(6): 4531–61. https://link.springer.com/10.1007/s00034-025-03410-4.

Haoxing, Zhuhai, and Control System. “No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title.” *Electronics* 10(17).

de Lope, Javier, and Manuel Graña. 2023. “An Ongoing Review of Speech Emotion Recognition.” *Neurocomputing* 528: 1–11.

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2023). 2023. “No Title Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation.”

Milton, A., S. Sharmy Roy, and S. Tamil Selvi. 2013. “SVM Scheme for Speech Emotion Recognition Using MFCC Feature.” *International Journal of Computer Applications* 69(9): 34–39. http://research.ijcaonline.org/volume69/number9/pxc3887667.pdf.

Neumann, Michael, and Ngoc Thang Vu. 2019. “Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech.” In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 7390–94.

Rifkin, R., & Klautau, A. 2004. “In Defense of One-Vs-All Classification".”

Sasongko, Sudi Mariyanto Al, Shofian Tsaury, Suthami Ariessaputra, and Syafaruddin Ch. 2023. “Mel Frequency Cepstral Coefficients (MFCC) Method and Multiple Adaline Neural Network Model for Speaker Identification.” *JOIV : International Journal on Informatics Visualization* 7(4): 2306. http://joiv.org/index.php/joiv/article/view/1376.

Satt, Aharon, Shai Rozenberg, and Ron Hoory. 2017. “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms.” In *Interspeech 2017*, ISCA: ISCA, 1089–93.

Abubakar, Abdulrahman et al. 2019. “Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition.” https://arxiv.org/abs/1907.06078.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20(3): 273–97.

Fayek, Haytham M. 2016. “Speech Processing for Machine Learning: Filter Banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What’s In-Between.” https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

George, Swapna Mol, and P Muhamed Ilyas. 2026. “Assessing the Effectiveness of Feature Normalization and Dataset Quality in

Speech Emotion Recognition Across Diverse Emotional and Linguistic Contexts." *Circuits, Systems, and Signal Processing* 45(6): 4531–61. https://link.springer.com/10.1007/s00034-025-03410-4.

Haoxing, Zhuhai, and Control System. "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title." *Electronics* 10(17).

de Lope, Javier, and Manuel Graña. 2023. "An Ongoing Review of Speech Emotion Recognition." *Neurocomputing* 528: 1–11.

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2023). 2023. "No Title Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation."

Milton, A., S. Sharmy Roy, and S. Tamil Selvi. 2013. "SVM Scheme for Speech Emotion Recognition Using MFCC Feature." *International Journal of Computer Applications* 69(9): 34–39. http://research.ijcaonline.org/volume69/number9/pxc3887667.pdf.

Neumann, Michael, and Ngoc Thang Vu. 2019. "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech." In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 7390–94.

Rifkin, R., & Klautau, A. 2004. "In Defense of One-Vs-All Classification"."

Sasongko, Sudi Mariyanto Al, Shofian Tsaury, Suthami Ariessaputra, and Syafaruddin Ch. 2023. "Mel Frequency Cepstral Coefficients (MFCC) Method and Multiple Adaline Neural Network Model for Speaker Identification." *JOIV : International Journal on Informatics Visualization* 7(4): 2306. http://joiv.org/index.php/joiv/article/view/1376.

Satt, Aharon, Shai Rozenberg, and Ron Hoory. 2017. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms." In *Interspeech 2017*, ISCA: ISCA, 1089–93.