Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

Pengenalan Emosi Ucapan Menggunakan Bidirectional Long Short Term Memory Pada Podcast Malaka

Kharisma Dwi Pratiwi*, Bambang Irawan, Otong Saeful Batchri
Teknik Informatika, Fakultas Teknik, Universitas Muhadi Setiabudi, Indonesia
Email: [1,*]kharismadwipratiwi@umus.ac.id, [2]bambangumus@gmail.com, [3]otongsaifulbahriumus@gmail.com
Email Penulis Korespondensi:kharismadwipratiwi@umus.ac.id*
Submitted: **12/01/2026**; Accepted: **13/02/2026**; Published: **31/03/2026**

Abstrak–Penelitian ini mengembangkan sistem pengenalan emosi ujaran berbahasa Indonesia menggunakan metode Bidirectional Long Short Term Memory (BiLSTM) dengan sumber data berupa ujaran alami dari podcast Malaka. Dataset diperoleh dari audio YouTube yang di konversi ke format WAV, disegmentasi, serta dilabeli emosi marah, netral, sedih dan senang berdasarkan acuan dataset CREMA-D. Tahapan prapemrosesan meliputi silce removal, normalisasi sinyal, dan augmentasi data. Ektrasi ciri dilakukan menggunakan Mel-Frequency Cepstral Coefficients (MFCC) beserta fitur delta dan delta-delta untuk mempresentasikan karakteristik spektral dan temporal sinyal suara. Model Bidirectional Long Short Term Memory dilatih dengan pembagian data 80% data latih, 10% data validasi, dan 10% data uji, serta dioptimasi dengan menggunakan algoritma adam. Hasil pengujian menunjukan bahwa model mencapai akurasi terbaik sebesar 80.32% dengan nilai pecision, recall, dan f1-score yang relatif seimbang pada seluruh kelas emosi. Hasil ini menunjukan bahwa Bidirectional Long Short Term Memory (BiLSTM) efektif dalam memodelkan dinamika temporal emosi pada ujaran podcast berbahasa indonesia.

Kata Kunci: Speech Emotion Rcognition; LSTM; BiLSTM; Emosi; Podcast

Abstract–This study develops an indonesian speech emotion recognition system using the Bidirectional Long Short Term Memory (BiLSTM) method with data sources in the form of natural speech from the Malaka podcast. The dataset was obtained from YouTube audio that was converted to WAV format, segmented, and labeled with angry, neutral, sad, and happy emotions based on the CREMA-D dataset reference. The preprocessing stages included slice removal, signal normalization, and data augmentation. Feature extraction was performed using Mel-Frequency Cepstral Coefficients (MFCC) along with delta and delta-delta features to represent the spectral and temporal characteristics of the voice signal. The Bidirectional Long Short Term Memory model was trained with a data split of 80% training data, 10% validation data, and 10% test data, and optimized using the Adam algorithm. The test results showed that the model achieved the best accuracy of 80.32% with relatively balanced precision, recall, and f1-score values across all emotion classes. These results show that Bidirectional Long Short Term Memory (LSTM) is effective in modeling the temporal dynamics of emotions in Indonesian podcast speech.

Keywords: Speech Emotion Recognition; LSTM; BiLSTM; Emotion; Podcast

1. PENDAHULUAN

Emosi merupakan aspek penting dalam komunikasi manusia dan sering diekspresikan melalui karakteristik vokal seperti intonasi, tempo, dan kualitas suara [1]. Dalam proses komunikasi emosi dapat dikenali melalui berbagai isyarat salah satunya adalah ucapan sehingga berkontribusi terhadap tercpitanya pemahaman timbal balik yang lebih baik. Dalam kajian psikologi dan linguistik emosi berkaitan erat dengan *affect* dan *mood* karena keduanya saling memengaruhi dalam membentuk kondisi psikologis individu [2]. Emosi dipahami sebagai keadaan mental yang bersifat dinamis dan muncul sebagai respon terhadap rangsangan internal maupun eksternal yang kemudian diekspresikan melalui perilaku atau reaksi tertentu dengan tingkat intensitas yang berbeda-beda [3]. Suara manusia membawa informasi berupa nada, kecepatan bicara, tekanan suara, dan intonasi yang mencerminkan kondisi emosional pembicara. Perubahan pada karakteristik akustik suara berkaitan erat dengan perubahan emosi seperti marah, sedih, atau senang [4].

Ucapan merupakan media komunikasi utama yang paling sering digunakan manusia untuk menyampaikan informasi, ide, maupun perasaan secara alami dan efisien [5]. Proses pengenalan emosi berdasarkan sinyal suara dikenal sebagai *speech emotion recognation*, yaitu bidang penelitian yang bertujuan memungkinkan sistem komputer mengenali dan mengklasifikasi emosi pembicara melalui karakteristik vokal. Penelitian *speech emotion recognation* berkembang karena memiliki potensi penerapan yang luas seperti pada interaksi manusia dan komputer, sistem percakapan cerdas, serta analisis kesehatan mental berbasis suara [6]. Dalam konteks kecerdasan buatan *speech emotion recognation* menggabungkan teknik *machine learning* dan pemrosesan sinyal untuk mengektrasi serta mempelajari pola emosi dari sinyal ucapan [7]. Perbedaan karakteristik suara antarindividu menjadikan proses ekstrasi fitur sebagai tahap yang sangat penting dalam sistem *speech emotion recognation* [8].

Salah satu ekstrasi fitur yang paling banyak digunakan adalah *mel frequency cepstral coefficients* yang di rancang berdasarkan skala *mel* agar sesuai dengan cara manusia mempersepsikan frekuensi suara dan terbukti

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 853
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

efektif dalam merepresentasikan karakteristik emosional pada sinyal ucapan. Informasi emosional dalam sinyal ucapan dapat direpresentasikan melalui fitur akustik dan prosodik [9]. Dalam sistem *speech emotion recognation* proses ekstrasi fitur bertujuan untuk menyederhanakan sinyal audio mentah tanpa menghilangkan karakteristik emosional yang penting. Pada pendekatan audio, karakteristik sinyal seperi amplitudo dan dinamika temporal memiliki pengaruh yang signifikan terhadap cara kerja pengenalan emosi [10]. Berbagai metode telah dikembangkan dalam pengenalan emosi berbasis ucapan, mulai dari algoritma konvensional hingga *deep learning*. Beberapa metode yang umum digunakan seperti *support vector machine, k-nearest neighbors, deep neural network, convulution neural network, recurrent neural network,* serta *long short term memory*. Pendekatan berbasis *deep learning* umumnya mampu menghasilkan tingkat akurasi yang lebih tinggi karena kemampuanya dalam mempelajari fitur yang kompleks meskipun memerlukan sumber daya yang lebih besar [11].

Penelitian terkait pengembangan *sistem speech emotion recognation* dalam berbahsa indonesia juga telah banyak dilakukan. Beberapa penelitian memanfaatkan korpus yang tersedia dan menerapkan berbagai algoritma klasifikasi *seperti support vector machine* dan *random forest* yang kemudian dikembangkan lebih lanjut menggunakan arsitekstur *long short term memory*. Disisi lain penelitian terbaru menunjukan bahwa *long short term memory* tetap mampu memberikan performa yang baik dalam pengenalan emosi berbasis audio apabila didukung oleh proses ektrasi fitur dan strategi pelatihan yang tepat [12]. Selain itu *long short term memory* direkomendasikan dalam berbagai penelitian karena kemampuanya mengatasi permasalahan *vanishing gradient* melalui mekanisme tiga gerbang utama yaitu *input gate, forget gate,* dan *output gate* yang memungkinkan jaringan mengontrol aliran informasi secara efektif dan menjaga kestabilan gradien selma proses pelatihan [13].

Namun pada percakapan alami seperti podcast, pehaman emosi tidak hanya dipengaruhi oleh konteks ucapan yang telah di ucapkan sebelumnya, tetapi juga oleh konteks ucapan setelahnya [14]. Emosi dalam komunikasi lisan bersifat berkala dan kontekstual, dimana makna emosional suatu ucapan seringkali baru dapat di pahami secara utuh setelah rangkaian ujaran berikutnya disampaikan, terutama pada percakapan spontan dan diskusi berdurasi panjang sehingga model *long short term memory* satu arah memproses data sekuensial hanya dari masa lalu ke masa depan (*forward direction*), sehingga representasi emosional yang di hasilkan sangat bergantung pada informasi sebelumnya [15]. Untuk mengatasi hal tersebut *bidirectional long short term memory* (BiLSTM) digunakan sebagai pengembangan dari *long short term memory* satu arah. *Bidirectional long short term memory* memproses sinyal ucapan melalui dua lapisan *long short term memory* yang berjalan secara pararel, yaitu lapisan *forward* yang mempelajari dependensi temporal dari masa depan dan lapisan *backward* yang mempelajari dependensi temporal dari masa depan ke masa lalu. Dengan mekanisme ini *bidirectional long short term memory* mampu memanfaatkan informasi kontekstual dari keseluruhan rangkaian ucapan, sehingga menghasilkan representasi fitur emosional yang lebih kaya dan komprehensif [16].

Pada penelitian terdahulu *bidirectional long short term memory* dilaporkan mampu menangkap konteks emosi secara lebih komprehensif karena proses pembelajaran model dilakukan dua arah serta menunjukan peningkatan performa dibandingkan *long short term memory* satu arah [17]. Pada penelitian yang membahas pendekatan berbasis *bidirectional long short term memory* untuk pengenalan emosi suara berbahasan indonesia menggunakan smote dengan dataset berbasis rekaman percakapan dari acara televisi *talk show* nasional dengan hasil akurasi sebesar 78% [18]. Kemudian penelitian yang dilakukan pada dataset *Interactive Emotional Dyadic Motion Capture* (IEMOCAP), *Berlin database Of Emotional Speech* (EMO-DB), *Ryerson Audio Visual Database of Emotional Speech and Song* (RAVDESS) mengguanakan pendekatan *bidirectional long short term memory* juga mendapatkan hasil akurasi 72.25% pada dataset IEMOCAP, 85.57% pada dataset EMO-DB dan 77.02% pada dataset RAVDESS [14]. Pada penelitian lain penggunaan *bidirectional long short term memory* juga menghasilkan hasil yang memuaskan pada dataset konvensional dengan akurasi 90.92% untuk dataset RAVDESS, 93% pada dataset EMO-DB, dan 92% pada dataset IEMOCAP [19].

Dari hasil penelitian *speech emotion recognation* kebanyakan masih menggunakan dataset konvensional berbasis aktor atau rekaman terkontrol seperti RAVDESS, EMO-DB, dan TESS[20]. Dataset semacam ini belum sepenuhnya merepresentasikan karakteristik percakapan alami dalam kehidupan sehari-hari yang bersifat spontan, berdurasi panjang, melibatkan banyak pembicara serta gangguan lingkungan. Kondisi tersebut berpotensi menyebabkan model *speech emotion recognation* memiliki keterbatasan generalisasi ketika diterapkan pada data alami. Dengan demikian terdapat celah penelitian dalam pengembangan sistem *speech emotion recognation* yang mampu mengenali emosi secara akurat pada data ucapan alami yang mencerminkan kondisi komunikasi manusia. Seiring dengan meningkatnya penggunaan media digital, *podcast* menjadi salah satu bentuk media audio yang berkembang pesat dan menyajikan percakapan yang bersifat alami dan spontan [3]. Berbeda dengan dataset berbasis aktor, percakapan dalam *podcast* merepresentasikan dinamika emosi yang lebih realistis sehingga

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 854
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

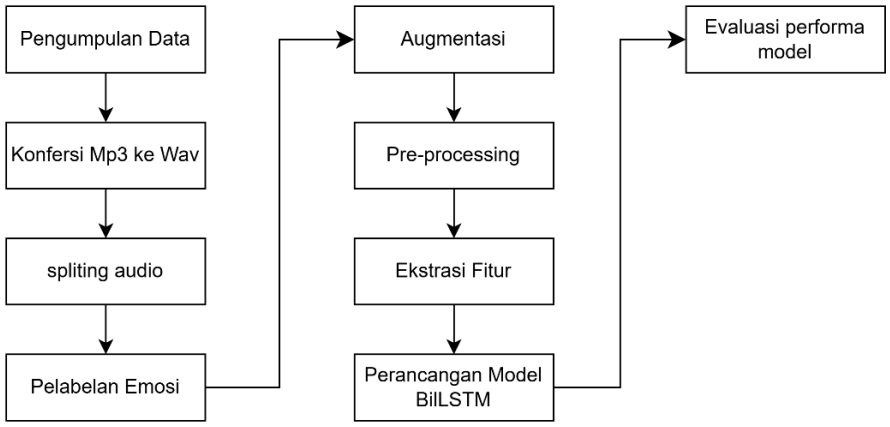
berpotensi digunakan sebagai sumber data dalam penelitian ini karena memuat percakapan spontan yang merepresentasikan emosi marah, sedih, senang, dan netral dalam konteks diskusi sehari-hari.

Berdasarkan permasalahan dan gap penelitian tersebut, penelitian ini bertujuan untuk mengembangkan dan mengevaluasi *sistem speech emotion recognation* berbasis fitur *mel frequency cepstral coefficient* dan arsitekstur *bidirectional long short term memory* pada data percakapan alami yang bersumber dari *podcast* malaka, serta menganalisis kemampuan model dalam mengenali emosi marah, senang, sedih, dan netral pada kondisi percakpan yang bersifat natural.

2. METODOLOGI PENELITIAN

2.1 Metode Penelitian

Metode penelitian ini menggunakan pendekatan eksperimental berbasis *deep learning* dengan tujuan mengembangkan dan mengevaluasi kinerja model *bidirectional speech emotion*. Pendekatan eksperimental ini dipilih karena penelitian ini melibatkan proses pelatihan model, pengujian performa, serta analisis hasil klasifikasi emosi berdasarkan data suara. Dalam pendekatan ini, peneliti merancang dan membangun model klasifikasi emosi kemudian mengamati pengaruh penerapan metode tertentu seperti menggunaan *mel frequency cepstral coefficients* dan arsitekstur *bidirectorial long short term memory* terhadap hasil pengenalan emosi ucapan. Melalui pendekatan kuantitatif ekperimental ini memungkinkan dilakukanya perbandingan kinerja model berdasarkan parameter dan konfigurasinya. Selain itu pendekatan ini memungkinkan analisis mendalam terhadap kemampuan model dalam mengenali pola emosi pada data audio yang bersifat natural dalam percakapan *podcast*. Dengan demikian penggunaan pendekatan kuantitatif ekperimental dinilai tepat untuk menjawab tujuan penelitian yaitu mengembangkan model *speech emotion recognation* yang mampu mengenali emosi ucapan secara akurat serta mengevaluasi kinerjanya berdasarkan *confulution matrix* yang jelas dan terukur. Tahap penelitian dapat dilihat pada Gambar 1.



Gambar 1. Blok Diagram Alur Penelitian

2.2 Pengumpulan Data

Pengumpulan data dilakukan dengan mengambil data audio dari tiga vidio *podcast* malaka yang berjudul "Ngga Ada yang Lebih Penting daripada Pendidikan dasar", "Isu Joki Viral Membongkar Cacat Pendidikan Indonesia", "Merawat ingatan Tentang Bangsa yang besar". Data audio kemudian diunduh dari platform YouTube dan disiapkan sebagai sumber utama untuk membangun dataset. Tahap ini penting karena kualitas dan variasi data akan snagat mempengaruhi performa model *speech emotion recognition*. Pemilihan *podcast* sebagai sumber data memberikan kelebihan berupa emosi yang muncul secara spontan atau alami, berbeda dengan dataset aktor seperti CREMA-D atau RAVDESS yang merupakan data berupa emosi terkontrol.

2.3 Konversi Format Audio

Pada tahap ini, file audio dalam format MP3 dikonversi ke format WAV. Format WAV dipilih karena memiliki kualitas tanpa kompresi sehingga mempertahankan informasi akustik secara penuh yang sangat penting dalam

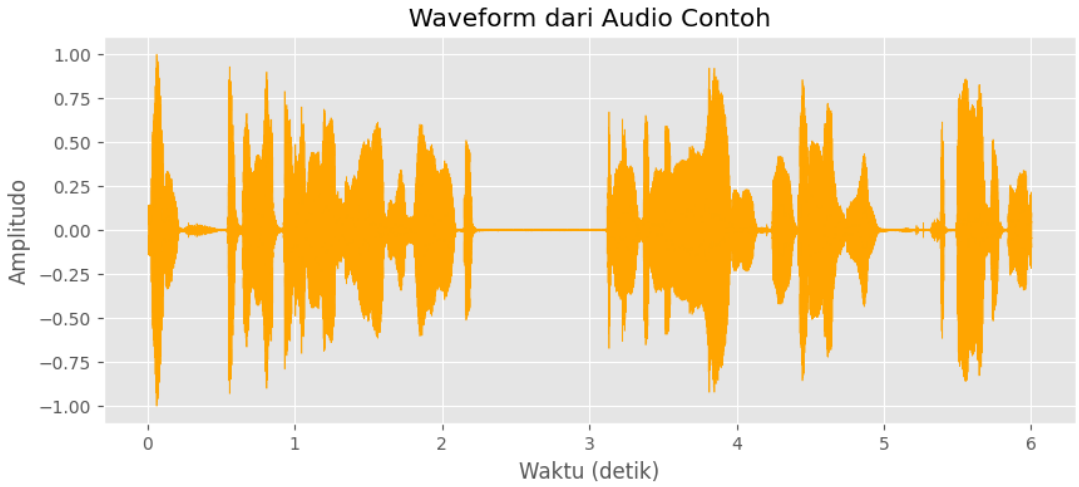
Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 855
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

ekstrasi fitur untuk *speech emotion recognition*. MP3 mengandung kompresi yang dapat menghilangkan detail penting dalam sinyal suara sehingga konversi diperlukan untuk menjaga kekuatan pemrosesan sinyal selanjutnya.

2.4 Segmentasi Audio

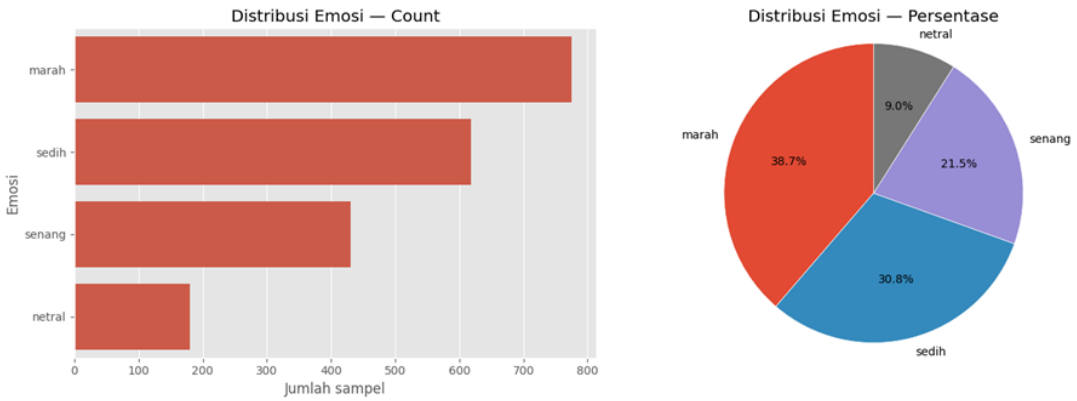
Audio *podcast* yang berdurasi panjang tidak dapat langsung digunakan sebagai input model. Oleh karena itu audio dibagi atau di *split* kedalam segmen-segmen yang lebih pendek. Pada penelitian ini pembagian tiap segmenya berdurasi selama 6 detik. Splitting dilakukan agar tiap sampel dapat lebih fokus pada suatu ekspresi emosi, serta memperbanyak jumlah data berupa data *augmentation* alami. Segmentasi juga memudahkan model dalam mendeteksi pola temporal yang sesuai dengan karakteristik emosi. Hasil spliting audio dapat dilihat pada Gambar 2.



Gambar 2. Contoh Hasil Waveform Audio

2.5 Pelabelan Emosi

Salah satu studi mengenai emosi ucapan dengan mengembangkan model klasifikasi ucapan menjadi 4 kelas emosi dengan menggunakan nada sebagai fiturnya. Pada penelitian ini setiap segmen audio diberi label emosi sesuai kategori yang digunakan yaitu marah, senang, sedih dan netral. Dalam penelitian ini peneliti menggunakan referensi pola emosi dari dataset CREMA-D sebagai pedoman karakter suara tiap emosi. Kemudian pengecekan setelah pelabelan dilakukan secara manual dengan mengacu pada *tone of voice*, intonasi, kecepatan bicara dan konteks percakapan. Proses ini sangan penting karena keberhasilan model sangat bergantung pada ketepatan label. Hasil pelabelan dan jumlah distribusi emosi dapat dilihat pada Gambar 3.



Gambar 3. Hasil Pelabelan Emosi

Gambar 3. terlihat jumlah emosi marah diangka 775 atau sekitar 38,7%, jumlah emosi sedih diangka 618 atau sekitar 30,8%, jumlah emosi senang diangka 431 atau sekitar 21,5% dan jumlah emosi netral diangka 180 atau sekitar 9%. Dari data hasil pelabelan pada Gambar 2. Dapat disimpulkan distribusi hasil pelabelan pada setiap

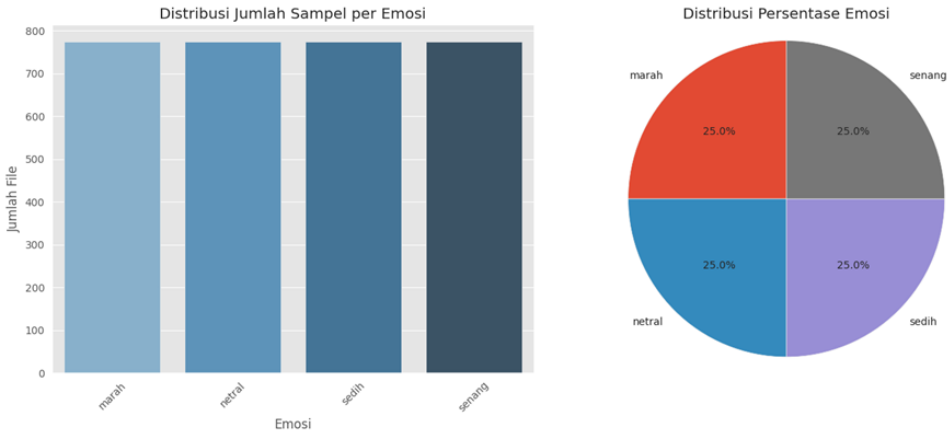
Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 856
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

emosi masih belum sama rata, maka pada proses berikutnya akan dilakukan augmentasi untuk menambahkan jumlah emosi pada setiap label.

2.6 Augmentasi

Penerapan teknik augmentasi data dalam studi pengenalan emosi ucapan menunjukan pentingnya teknik-teknik ini untuk pengenalan suara yang efektif diberbagai bidang penelitian. Dengan beberapa tantangan yang diidentifikasi pada dataset suara yang ada, terutama kurangnya dataset yang memadai dan ketidak seimbangan kelas yang berdampak besar pada kinerja pengklasifikasian, oleh karena itu kebutuhan untuk meningkatkan sampel data dan menyeimbangkan distribusi kelas sangat penting untuk meningkatkan sistem pengenalan [21]. Dataset hasil augmentasi dapat dilihat pada Gambar 4.



Gambar 4. Hasil Augmentasi Emosi

Gambar 4. Terlihat jumlah emosi pada tiap distribusi emosi sudah sama rata dengan jumlah masing-masing emosi pada distribusi emosi berjumlah 750 atau sekitar 25%.

2.7 Pre-processing

Pada tahap ini merupakan sebuah tahapan untuk melakukan normalisasi terhadap suara yang didapatkan pada tahap sebelumya. Tahapan ini mencangkup pemrosesan sinyal audio untuk meningkatkan fitur. Proses yang dilakukan diantaranya yaitu *silence removal*. *Silence removal* berfungsi menghapus bagian yang tidak memiliki suara agar model tidak belajar pola yang tidak relevan. Kemudian ada *noise reduction* yang berfungsi untuk mengurangi kebisinga latar karena *podcast* direkam bukan dalam ruang studio. Selanjutnya ada normalisasi amplitudo yang berfungsi untuk menyamakan intesitas suara agar perbedaan volume tidak memengaruhi ekstrasi fitur. Terakhir pada proses *pre-processing* terdapat *framing* dan *windowing*, proses tersebut digunakan untuk memproses input suara dipotong menjadi frame-frame dengan durasi yang lebih pendek dan bertujuan untuk memperoleh sampel sinyal yang tepat. *Pre-processing* memastikan sinya siap untuk diesktrasi fitur dan menjaga kualitas data di seluruh pipeline.

2.8 Ekstrasi Fitur

Pada tahap ini dilakukan proses ektrasi fitur menggunakan *mel frequency cepstral coefficient* yaitu metode yang banyak digunakan dalam pengenalan emosi berbasis suara karenaaa meniru cara kerja sistem pendengaran manusia [7] .Pada penelitian ini alur ekstrasi pada *mel frequency cepstral coefficient* dilakukan dengan memotong sinyal menjadi frame kecil (*frame lenght dn hop lenght*).

2.9 Perancangan Model BiLSTM

Bidirectional long short term memory adalah arsitekstur pengembangan dari *long short term memory* yang mampu mempelajari ketergantungan jangka panjang (*long term dependencies*) dalam data sekuensial seperti sinyal audio dam setiap unit *bidirectional long short term memory* memiliki enam komponen utama yaitu *input gate*, *forget gate*, *candidate cell state*, *cell state update*, *output gate* dan *hidden state* sebagaimana *long short term memory* konvensional. Arsitekstur BiLSTM memiliki dua lapisan yang bekerja secara pararel dan berlawanan arah

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 857
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

yaitu *forward long short term memory* t = 1 → T dan arah *backward long short term memory* t = T → 1 dengan d ∈{f,b} menunjukan arah *forward* dan *backward* pada arsitekstur BiLSTM [22]. *Input gate* digunakan untuk mengatur seberapa banyak informasi baru dari input saat ini (x_t) yang akan disimpan ke dalam *cell state* pada BiLSTM perhitungan dilakukan dua kali arah maju dan arah mundur sebagai contoh ($i_t^{(f)}$) menunjukan gate untuk arah maju dan $i_t^{(d)}$ untuk arah mundur [23] (Persamaan 1). Pada penelitian ini, *input gate* membantu model menentukan seberapa besar pengaruh fitur audio pada frame tertenttu yang harus disimpan dalam memori jangka panjang untuk mendeteksi ekspresi emosi seperti marah, senang, sedih, atau senang.

$$i_t^{(d)} = \sigma\left(w_i^{(d)}\left[h_{t-1}^{(d)}, x_t\right] + b_i^{(d)}\right) \quad (1)$$

Forget gate untuk mengontrol informasi dari *cell state* sebelumnya (c_{t-1}) yang perlu dilupakan (Persamaan 3). Pada *gate* ini model menghapus informasi yang tidak relevan seperti pada bagian hening atau *noise*, sehingga model fokus pada bagian sinyal yang memiliki makna emosional dan dilakukan secara pararel.

$$f_t^{(d)} = \sigma\left(w_f^{(d)}\left[h_{t-1}^{(d)}, x_t\right] + b_f^{(d)}\right) \quad (2)$$

Candidate cell state menghasilkan nilai baru yang akan memperbarui *cell state* [18] (Persamaan 3). Bagian ini memproses informasi baru berdasarkan input suara yang ada, seperti perubahan nada atau tekanan suara yang berperan penting dalam membentuk pola emosi.

$$c_{\sim t}^{(d)} = \tanh\left(w_c^{(d)}\left[h_{t-1}^{(d)}, x_t\right] + b_c^{(d)}\right) \quad (3)$$

Update cell state menentukan *cell state* baru berdasarkan kombinasi *forgate gate* dan input (Persamaan 4). Pada proses ini model mengingat urutan perubahan suara dalam waktu panjangg misalnya fluktasi energi suara saat seseorang berbicara marah atau sedih sehingga *long short term memory* menyimpan pola emosi antar frame audio bukan frame tunggal.

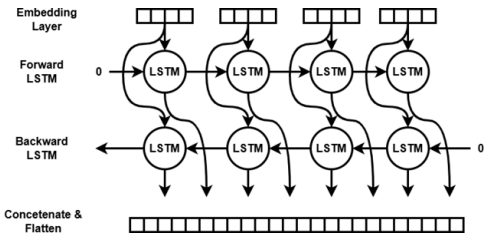
$$c_t^{(d)} = f_t^{(d)} \times c_{t-1}^{(d)} + i_t^{(d)} \times c_{\sim t}^{(d)} \quad (4)$$

*Output gat*e mengontrol bagian dari *cell state* yang digunakan untuk menghitung *hidden state* ($h_t^{(b)}$) [24] (Persamaan 5). Gate ini mengatur informasi dari memori yang akan di keluarkan pada waktu saat ini untuk proses klasifikasi emosi sehingga yang di hasilkan hanya informasi emosi yang paling relevan pada setiap waktu.

$$o_t^{(d)} = \sigma\left(w_o^{(d)}\left[h_{t-1}^{(d)}, x_t\right] + b_o^{(d)}\right) \quad (5)$$

Hidden state merupakan *output gate* dari unit *long short term memory* pada waktu ke-t yang akan diteruskan ke langkah berikutnya (Persamaan 6). Dalam penelitian ini *hidden state* menjadi representasi akhir dari kondisi emosi suara yang selanjutnya diteruskan ke lapisan *dense* untuk klasifikasi menjadi empat kategori emosi yaitu marah, senang, sedih, netral.

$$h_t = [h_t^{(f)} \,||\, h_t^{(b)}] \quad (6)$$



Gambar 5. Arsitekstur BiLSTM [18]

2.10 Evaluasi Performa Model

Setelah pelatihan selesai, model dievaluasi untuk mengukur kemampuan dalam mengenali emosi. Evaluasi dilakukan menggunakan akurasi, presisi, r*ecall*, *f1-score* dan *confusion matrix*.

3. HASIL DAN PEMBAHASAN

3.1 Evaluasi Hasil

3.1.1 Ekstrasi Audio

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 858
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

Kombinasi fitur *mel frequency cepstral coefficient*, *zero crossing rate*, dan *root mean square* mampu merepresentasikan karakteristik emosional sinyal ucapan secara numerik. *Mel frequency cepstral coefficient* menangkap informasi spektral dan intonasi suara, sedangkan *zero crossing rate* dan *root mean square* merepresentasikan perubahan frekuensi serta energi sinyal ucapan. Penggunaan kombinasi fitur ini memungkinkan model mempelajari pola temporal dan karakteristik emosional suara secara bertahap yang ditunjukan oleh peningkatan akurasi selama proses pelatihan hingga mencapai performa terbaik. Representasi numerik ini selanjutnya digunakan sebagai input utama bagi model *bidirectional long short term memory* untuk melakukan klasifikasi emosi ucapan dan dapat dilihat pada Tabel 1.

Tabel 1. Hasil Ekstrasi Audio

zcr	rms	mfcc	path	label
0.083789	0.093240	-14.425267	/content/drive/MyDrive/youtubeset/audiosplit/dataset18_part437.wav	senang
0.071883	0.054813	-18.887260	/content/drive/MyDrive/youtubeset/audiosplit/dataset18_part537.wav	sedih
0.076281	0.080695	-12.200109	/content/drive/MyDrive/youtubeset/audiosplit/dataset10_part5.wav	marah
0.078598	0.051617	-19.327497	/content/drive/MyDrive/youtubeset/audiosplit/dataset19_part795.wav	netral

3.1.2 Dataset Untuk Pelatihan Model

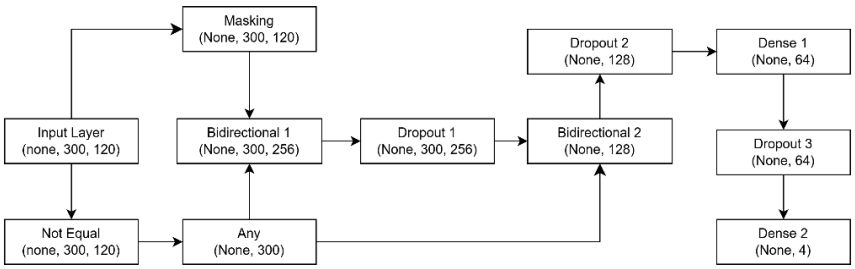
Hasil Dataset yang digunakan berasal dari potongan audio *podcast* malaka yang telah melalui proses konversi format, pemotongan audio, pelabelan emosi, augmentasi data, serta nomalisasi panjang sinyal *pading*. Dataset dibagi menjadi data *train* sebanyak 80% atau 2480 audio, data *validation* sebanyak 10% atau 310 audio, dan data *test* sebanyak 10% atau 310 audio dimana masing-masing sampel memiliki 300 frame waktu dan 120 fitur perframe. Representasi ini sesuai dengan karakteristik metode *bidirectional long short term memory* yang memanfaatkan informasi temporal untuk mempelajari pola emosi pada sinyal ucapan. Hasil ekstrasi dapat dilihat pada Tabel 2.

Tabel 2. Pembagian Dataset

Subset	Jumlah Sampel	Proporsi	Shape Input	Shape Target
Train	2480	80%	2480, 300, 120	2480, 4
Val	310	10%	310, 300, 120	310, 4
Test	310	10%	310, 300, 120	310, 4

3.1.3 Hasil Pelatihan Model BiLSTM

Hasil pelatihan menunjukan bahwa model *bidirectional long short-term memory* mampu mempelajari pola temporal dua arah dari fitur audio yang di ekstrasi menggunakan *mel-frequency cepstral coefficient*. Kombinasi dua lapisan *bidirectional long short-term memory*, lapisan *drop out*, dan lapisan *dense* menghasilkan sistem klasifikasi emosi yang stabil dengan kemampuan generalisasi yang cukup baik dan dapat dilihat pada Gambar 6.



Gambar 6. Blok Diagram Arsitekstur Model BiLSTM

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 859
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

Model *bidirectional long short-term memory* dalam penelitian ini dikembangkan menggunakan *library* keras untuk pembelajaran mesin dan kecerdasan buatan. Model *bidirectional long short term memory* memiliki dua lapisan *bidirectional long short-term memory* dengan masing-masing 128 dan 64 unit *neuron*, serta beberapa lapisan *dense* dan *dropout* dengan nilai 0,3 untuk mengurangi *overfitting*. Arsitektur model ini dapat dilihat pada Tabel 3 *log summary*.

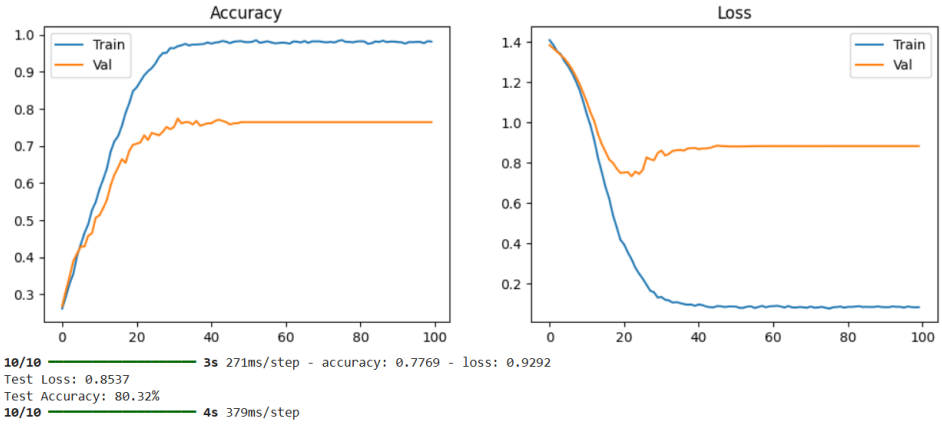
Tabel 3. *Log Summary* dari Arsitekstur BiLSTM

Layer	Output Shape	Param
Input layer	(None, 300, 120)	0
Not equal	(None, 300, 120)	0
Masking	(None, 300, 120)	0
Any	(None, 300)	0
bidirectional 1	(None, 300, 256)	254,976
Dropout 1	(None, 300, 256)	0
bidirectional 2	(None, 128)	164,352
dropout 2	(None, 128)	0
dense 1	(None, 64)	8,256
dropout 3	(None, 64)	0
dense 2	(None, 4)	260
	Total Params	427,844
	Trainable Params	427,844
	Non-trainable Params	0

Model *bidirectional long short-term memory* kemudian dilatih menggunakan data latih hasil ekstrasi fitur audio berbasis MFCC, ZCR, dan RMS untuk melakukan klasifikasi terhadap empat kelas emosi yaitu marah, netral, sedih dan senang.

3.1.4 Nilai Akurasi Model BiLSTM

Model *bidirectional long short term memory* yang dibangun pada penelitian ini di latih menggunakan fitur *mel frequency cepstral coefficient* yang dikombinaksikan dengan fitur turunan *delta-delta*. Model di kompilasi menggunakan *optimizer adam* dengan *learning rate* 0.0001 menggunakan 100 *epoch*.



Gambar 7. Grafik Akurasi dan Loss

Pada Gambar 7 akurasi data latih meningkat tajam sejak epoch awal hingga mendekati 100% pada sekitar *epoch* ke-30, sedangkan akurasi data validasi berhenti meningkat pada kisaran 77% sekitar *epoch* ke-25. *Loss* data latih menurun secara konsisten, sementara *loss* data validasi cenderung stagnan, kemudian meningkat dan stabil pada nilai yang lebih tinggi. Kondisi ini secara jelas menunjukan bahwa model mengalami *overfitting*, namun dengan nilai *test accuracy* di atas 80% model masih dapat dikatakan memiliki performa yang cukup baik.

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 860
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

Tabel 4. Hasil Training

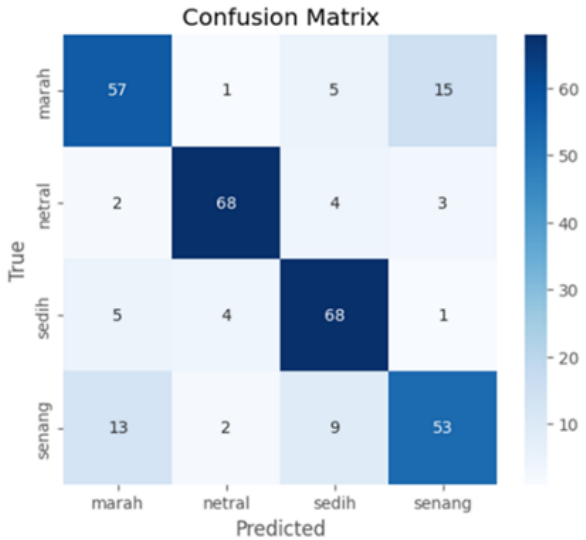
	Epoch	Train Accuracy	Train Loss	Val Accuracy	Val Loss
Best Epoch	23	0.8876	0.3243	0.7290	0.7331
Worst Epoch	1	0.2566	1.4109	0.2677	1.3840

Berdasarkan pada Tabel 4. model mencapai performa terbaik pada epoch ke-23 karena model menghasilkan *validation accuracy* tertinggi sebesar 0.7290 dan *validation loss* terendah sebesar 0.7331. Pada *epoch* yang sama, model memperoleh nilai *train accuracy* sebesar 0.8876 dan *train loss* sebesar 0.3243 menunjukan bahwa model telah mempelajari data latih secara optimal dan masih mempertahankan kemampuan generalisasi terhadap data validasi. Sebaliknya, model menunjukan performa terburuk pada *epoch* ke-1 karena model hanya menghasilkan *train accuracy* sebesar 0,2566 dan *validation accuracy* sebesar 0.2677 pada tahap awal pelatihan. Selain itu model memeroleh nilai *train loss* sebesar 1.4109 dan *validation loss* sebesar 1.3840 menunjukan model belum mampu mengenali pola data secara efektif.

Tabel 5. Hasil Evaluasi Model Data Uji

Label	Precision	Recall	F1-Score	Support
marah	0.74	0.78	0.76	78
netral	0.88	0.90	0.89	77
sedih	0.81	0.82	0.82	78
senang	0.77	0.71	0.74	77
accuracy			0.80	310
macro avg	0.80	0.80	0.80	310
weighted	0.80	0.80	0.80	310

Confusion matix ini memberikan informasi mengenai jumlah prediksi yang benar dan salah pada setipa kelas, sehingga dapat menunjukan pola kesalahan klasifikasi yang dilakukan oleh model. Visual *confusion matrix* hasil klasifikasi emosi pada data uji ditunjukan pada Gambar 8.



Gambar 8. Confusion Matrix Hasil Klasifikasi Emosi

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 861
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9432

Tabel 5 pada emosi netral menunjukan performa terbaik *f1-score* sebesar 0,89 yang mengindikasikan bahwa karakteristik emosi netral lebih konsisten dan mudah dikenali oleh model. Emosi sedih juga menunjukan performa yang baik dengan *f1-score* sebesar 0.82 mengindikasikan keberhasilan model dalam menangkap pola ucapan emosi negatif. Emosi marah memiliki *f1-score* sebesar 0.76 dan senag memperoleh *f1-score* 0.74. aberdasarkan Gambar 8, meskupun akurasi keseluruhan mencapai 0.80, beberapa kelas emosi masih mengalami kesalahan klasifikasi, terutama yang diklasifikasikan sebagai marah dan senang, yang menunjukan bahwa model masih kesulitan untuk secara konsisten membedakan kelas emosi tertentu.

3.1 Pembahasan

Arsitekstur *bidirectional long short term memory* degan fitur *mel-frequency cepstral coefficient* mampu menangkap pola temporal emosi dalam percakapan *podcast* berbahasa indonesia dengan cukup efektif. Hasil evaluasi perkelas emosi juga menunjukan bahwa emosi netral memiliki performa terbaik dengan nilai *precision* 0.88, *recall* 0,90, dan *f1-score* 0.89 yang mengindikasikan bahwa karakteristik akustik emosi netral lebih konsisten dan mudah dikenali oleh model. Namun pada hasil *confusion matrik* memperlihatkan bahwa kesalahan klasifikasi paling sering terjadi antara emosi senang-marah dan senang-sedih yang menandakan tantangan dalam membedakan emosi dengan dinamika energi dan intonasi yang saling tumpang tindih, akan tetapi distribusi antar kelas relatif seimbang menunjukan model tidak bias terhadap satu sama lain. Berdasarkan hasil pengujian, model *bidirectional long short term memory* mampu mencapai akurasi pengujian sebesar 80,32% dan *best epoch* ada pada *epoch* ke 23 dan *worst epoch* ada pada *epoch* ke 1. Meskipun pada pengujian model masih mengalami *overfitting* pada beberapa kelas emosi, hasil ini selaras dengan penelitian terdahulu [14] bahwa model *bidirectional long short term memory* memiliki kemampuan yang cukup baik dalam mengenali emosi sinyal ucapan meskipun di uji pada data *podcast* yang memiliki karakteristik percakapan yang spontan dan natural sebagai data percakapan alami pada bahasa dengan sumber data terbatas.

4. KESIMPULAN

Penggunaan dataset audio dari *podcast* malaka memberikan tantangan tersendiri dibandingkan dataset seperti RAVDESS, CREMA-D atau dataset lainya. Audio *podcast* memiliki karakteristik variasi emosi yang lebih natural, adanya *noise* latar, perubahan jarak suara dan variasi gaya bicara yang mempengaruhi kestabilan fitur. Penelitian ini berhasil mengembangkan model pengenalan emosi ucapan berbahasa indonesia menggunakan *bidirectional long short term memory* dengan memanfaatkan fitur *mel-frequency cepstral coefficient* beserta turuanya yaitu *delta* dan *delta-delta*. Hasil pengujian menunjukan bahwa model mampu mengenali empat kelas emosi dengan performa yang cukup baik, ditunjukan oleh akurasi pengujian sebesar 80.32% serta nillai *precision, recall*, dan *f1-score* yang relatif seimbang pada seluruh emosi. Hasil training model pada penelitian ini diperoleh *best epoch* pada *epoch* ke 23 dengan nilai *training accuracy* 88.76%, *train loss* sebesar 32,43%, val accuracy 73% dan *val loss* sebesar 73,31% dan diperoleh *worst epoch* pada epoch ke 1 degan nilai train accuracy 25.66%, train loss 1.41%, val accuracy 26,77% dan val loss 1.38%. Penelitian ini mengindikasikan bahwa kombinasi *pre-processing, silence removal* dan representasi fitur *mel-frequency cepstral coefficient* efektif dalam meningkatkan kemampuan model dalam menangkap karakteristik siyal emosi. Perbedaan antara akurasi pelatihan dan akurasi validasi menunjukan adanya kecenderungan *overfitting*. Penelitian ini memiliki keterbatasan pada jumlah dan keberagaman data yang masih terbatas sehingga perlu adanya pengembangan lebih lanjut pada aspek regulasi dan perluasan data berbahasa indonesia. Penelitian selanjutnya diharapkan dapat memperluas sumber data ke berbagai podcast atau percakapan spontan berbahasa indonesia serta mengekplorasi arsitektur model yang lebih kompleks untuk meningkatkan generalisasi sitem.

REFERENCES

[1] J. Nesi, E. H. Telzer, and M. J. Prinstein, *Handbook of Adolescent Digital Media Use and Mental Health*. 2022. doi: 10.1017/9781108976237.

[2] F. KASYIDI, R. ILYAS, and N. M. ANNISA, “Peningkatan Kemampuan Pengenalan Emosi Melalui Suara dalam Bahasa Indonesia,” *MIND J.*, vol. 6, no. 2, pp. 194–204, Dec. 2021, doi: 10.26760/mindjournal.v6i2.194-204.

[3] T. B. Putri, S. Saidah, B. Hidayat, F. Qothrunnada, T. Telekomunikasi, and U. Telkom, “Deteksi Emosi Berdasarkan Sinyal Suara Manusia Menggunakan Discrete Wavelet Transform (DWT) Dengan Klasifikasi Support Vector Machine (SVM) antar manusia tidak selalu terjadi dengan baik , ada beberapa faktor dari interaksi yang dapat berekspresi atau,” vol. 3, no. 1, pp. 1–10, 2023.

[4] S. Helmiyah, A. Fadlil, A. Yudhana, A. Dahlan, and P. Studi Teknik Elektro, “Pengenalan Pola Emosi Manusia Berdasarkan Ucapan Menggunakan Ekstraksi Fitur Mel-Frequency Cepstral Coefficients (MFCC) Speech Based Emotion Pattern Recognition Using Mel-Frequency Cepstral Coefficients (MFCC) Feature Extraction,” *Cogito Smart*

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 862
This Journal is licensed under a Creative Commons Attribution 4.0 International License

J., vol. 4, no. 2, p. 372, 2018.
[5] F. J. Tanudjaja, E. Y. Puspaningrum, and Y. V. Via, "Klasifikasi Jenis Emosi Melalui Ucapan Menggunakan Metode Convolutional Neural Network," *Teknologi*, vol. 13, no. 2, pp. 1–11, 2023, doi: 10.26594/teknologi.v13i2.3740.
[6] H. K. Bhuyan, B. Brahma, N. K. Kamila, S. Peram, B. Leelambika, and A. Sahu, "Mel-Spectrograms Based LSTM Model for Speech Emotion Recognition," *Trait. du Signal*, vol. 42, no. 3, pp. 1353–1365, 2025, doi: 10.18280/ts.420312.
[7] R. A. Nawasta, N. H. Cahyana, and H. Heriyanto, "Implementation of Mel-Frequency Cepstral Coefficient as Feature Extraction using K-Nearest Neighbor for Emotion Detection Based on Voice Intonation," *Telematika*, vol. 20, no. 1, p. 51, 2023, doi: 10.31315/telematika.v20i1.9518.
[8] N. Aini Lailla Asri, R. Ibnu Adam, and B. Arif Dermawan, "Speech Recognition Untuk Klasifikasi Pengucapan Nama Hewan Dalam Bahasa Sunda Menggunakan Metode Long-Short Term Memory," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1242–1247, 2023, doi: 10.36040/jati.v7i2.6744.
[9] N. D. Pah, *PEMROSESAN SINYAL DIGITAL*, Edisi Pert. Yogyakarta: Graha Ilmu, 2018.
[10] C. Roy *et al.*, "Stacked convolutional neural network for emotion recognition using multi feature speech analysis," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-28766-0.
[11] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.
[12] Q. Hu, Y. Peng, and Z. Zheng, "A deep learning framework for gender sensitive speech emotion recognition based on MFCC feature selection and SHAP analysis," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-14016-w.
[13] H. K. Bhuyan, B. Brahma, B. Leelambika, and S. Peram, "Machine Translated by Google Pemrosesan Sinyal Model LSTM Berbasis Spektrogram Mel untuk Pengenalan Emosi Ucapan Machine Translated by Google," vol. 42, no. 3, pp. 1353–1365, 2025.
[14] N. Senthilkumar, S. Karpakam, M. Gayathri Devi, R. Balakumaresan, and P. Dhilipkumar, "Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks," *Mater. Today Proc.*, vol. 57, pp. 2180–2184, 2022, doi: 10.1016/j.matpr.2021.12.246.
[15] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Syst. Appl.*, vol. 173, no. September 2019, p. 114683, 2021, doi: 10.1016/j.eswa.2021.114683.
[16] M. V. Subbarao, S. K. Terlapu, and P. S. R. Chowdary, "Emotion Recognition using BiLSTM Classifier," *Proc. - 2022 Int. Conf. Comput. Commun. Power Technol. IC3P 2022*, pp. 195–198, 2022, doi: 10.1109/IC3P52835.2022.00048.
[17] P. L. Seabe, C. R. B. Moutsinga, and E. Pindza, "Sentiment-driven cryptocurrency forecasting: analyzing LSTM, GRU, Bi-LSTM, and temporal attention model (TAM)," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, pp. 1–27, 2025, doi: 10.1007/s13278-025-01463-6.
[18] N. Nur Shabrina, F. Kasyidi, and R. Ilyas, "A BiLSTM-Based Approach For Speech Emotion Recognition In Conversational Indonesian Audio using SMOTE," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3173–3187, 2025, doi: 10.52436/1.jutif.2025.6.5.5183.
[19] F. Harby, M. Alohali, A. Thaljaoui, and A. S. Talaat, "Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition," *Comput. Mater. Contin.*, vol. 78, no. 2, pp. 2689–2719, 2024, doi: 10.32604/cmc.2024.046623.
[20] S. R. Livingstone and F. A. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. 2018.
[21] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," *Electron.*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223795.
[22] B. Jang, M. Kim, G. Harerimana, S. U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, 2020, doi: 10.3390/app10175841.
[23] N. Nosrati and Z. Navabi, "Analysis and Enhancement of Resilience for LSTM Accelerators Using Residue-Based CEDs," *IEEE Access*, vol. 12, no. February, pp. 52851–52866, 2024, doi: 10.1109/ACCESS.2024.3386431.
[24] L. Khan, A. Qazi, H. T. Chang, M. Alhajlah, and A. Mahmood, "Empowering Urdu sentiment analysis: an attention-based stacked CNN-Bi-LSTM DNN with multilingual BERT," *Complex Intell. Syst.*, vol. 11, no. 1, pp. 1–14, 2025, doi: 10.1007/s40747-024-01631-9.

Copyright © 2026 **Kharisma Dwi Pratiwi**, Page 863
This Journal is licensed under a Creative Commons Attribution 4.0 International License