BAB II

LANDASAN TEORI

2.1. *Speech Emotion Recognition*

Speech Emotion Recognition (SER) merupakan salah satu bidang penelitian dalam kecerdasan buatan yang bertujuan untuk mengenali kondisi emosional pembicara berdasarkan sinyal ucapan. Emosi pada ucapan dapat direpresentasikan melalui berbagai karakteristik akustik seperti *pitch,* energi, intonasi, ritme, dan dinamika temporal suara. Informasi tersebut memungkinkan sistem komputer untuk mengidentifikasi keadaan emosional pembicara secara otomatis. Perkembangan metode *deep learning* dalam beberapa tahun terakhir telah meningkatkan performa sistem SER karena mampu mempelajari representasi emosi dari sinyal audio secara lebih efektif dibandingkan metode konvensional. Berbagai pendekatan seperti *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), dan *Attention Mechanism* banyak digunakan untuk mengekstraksi pola emosional yang kompleks dari data ucapan (Abbaschian dkk., 2021).

Perkembangan metode *deep learning* mendorong peningkatan performa sistem SER karena mampu mempelajari representasi emosi secara otomatis dari data audio. Berbagai arsitektur *deep learning* banyak digunakan untuk mengekstraksi pola emosional yang kompleks dari sinyal ucapan. Tantangan seperti keterbatasan dataset, variasi cara berbicara, *noise* lingkungan, serta perbedaan persepsi emosi antar individu masih menjadi hambatan utama dalam

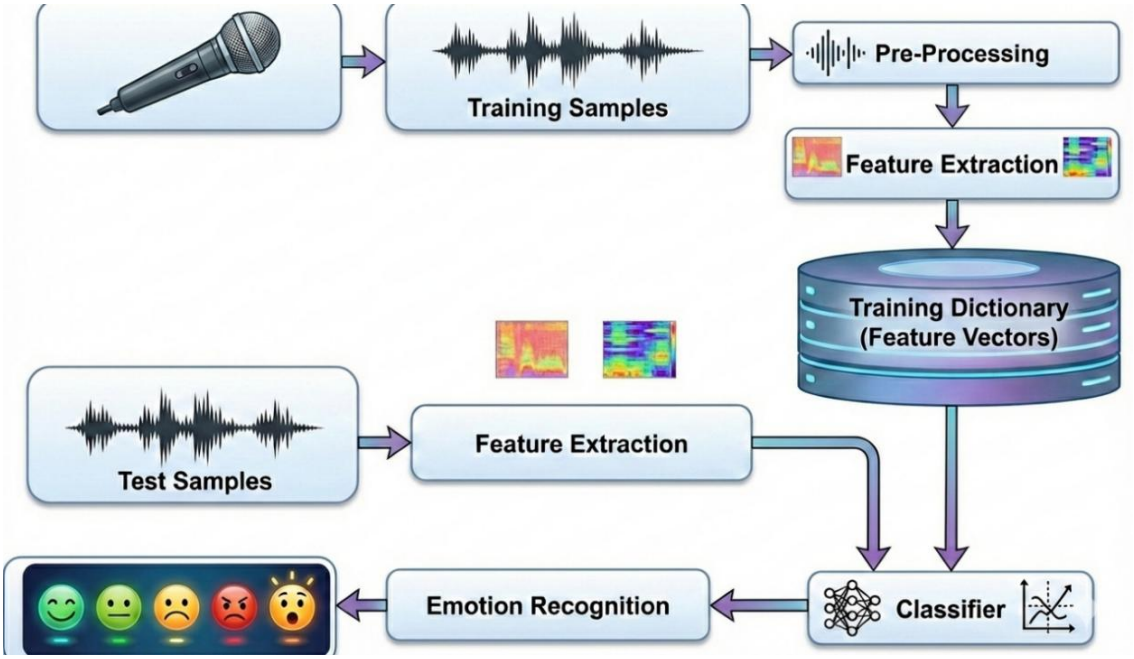
pengembangan sistem SER (Liu dkk., 2023).

Sistem SER memiliki alur terstruktur yang dimulai dari input audio hingga menghasilkan prediksi emosi. Alur tersebut meliputi pengambilan sinyal suara, pra-pemrosesan audio, ekstraksi fitur akustik, proses pembelajaran menggunakan *machine learning* atau *deep learning*, serta evaluasi performa model. Beberapa penelitian juga menerapkan teknik augmentasi data untuk meningkatkan variasi data pelatihan dan memperbaiki kemampuan generalisasi model (Barhoumi & Benayed, 2025).

Arsitektur SER umumnya terdiri dari tahap pelatihan (*training*) dan tahap pengujian (*testing*). Tahap pelatihan diawali dengan proses pra-pemrosesan sinyal suara untuk meningkatkan kualitas data, seperti normalisasi dan penghilangan bagian hening (*silence trimming*). Sinyal audio kemudian diekstraksi menjadi fitur-fitur akustik yang merepresentasikan karakteristik emosional suara. Fitur tersebut kemudian digunakan sebagai masukan bagi model klasifikasi untuk mempelajari pola emosi dari data pelatihan (Kamble dkk., 2015).

Tahap pengujian memproses sinyal suara baru menggunakan tahapan ekstraksi fitur yang sama seperti pada data pelatihan. Fitur yang dihasilkan kemudian dimasukkan ke dalam model klasifikasi yang telah dilatih sebelumnya untuk menentukan kategori emosi pembicara. Hasil akhir sistem SER berupa label emosi tertentu, seperti senang, netral, marah, atau sedih. Alur tersebut memungkinkan sistem SER melakukan pengenalan emosi secara sistematis berdasarkan karakteristik suara pembicara (Kamble dkk., 2015). Arsitektur umum

sistem SER ditunjukkan pada **Gambar 2.1**.



Gambar 2.1. Arsitektur SER (Dimodifikasi dari (Kamble dkk., 2015))

2.2. Sinyal dan Pengolahan Audio Digital

Sinyal audio digital merupakan representasi sinyal suara dalam bentuk data diskrit yang diperoleh melalui proses digitalisasi sinyal analog sehingga dapat diproses oleh komputer. Sinyal audio dalam bidang pengolahan ucapan (*speech processing*), mengandung berbagai informasi akustik seperti amplitudo, frekuensi, energi, dan karakteristik temporal yang dapat digunakan untuk menganalisis ucapan manusia, termasuk pengenalan emosi pada sistem SER (Seema, 2020).

2.2.1. *Sampling Audio*

Suara manusia dihasilkan dalam bentuk sinyal kontinu (*continuous signal*) yang berubah terhadap waktu. Proses digitalisasi memerlukan *sampling*, yaitu pengambilan nilai amplitudo sinyal pada interval waktu tertentu sehingga

menghasilkan sinyal diskrit. Frekuensi *sampling* menunjukkan jumlah sampel yang diambil setiap detik dan dinyatakan dalam satuan *Hertz* (Hz). Frekuensi *sampling* yang lebih tinggi menghasilkan representasi sinyal digital yang semakin mendekati sinyal asli, tetapi juga meningkatkan kebutuhan komputasi (Ambaye, 2020).

Proses *sampling* dapat dinyatakan sebagai Persamaan (2.1) (Ambaye, 2020):

$$x[n] = x(nT_s) \tag{2.1}$$

Keterangan:
$x[n]$ adalah sinyal diskrit hasil *sampling*,
x(t) adalah sinyal kontinu,
T_s adalah periode *sampling*,
n adalah indeks sampel.

Hubungan antara periode *sampling* dan frekuensi *sampling* dirumuskan sebagai Persamaan (2.2) (Ambaye, 2020):

$$f_s = \frac{1}{T_s} \tag{2.2}$$

Keterangan:
f_s adalah frekuensi *sampling,*
T_s adalah periode *sampling*,

2.2.2 *Domain* Waktu dan *Domain* Frekuensi

Analisis sinyal audio digital dapat dilakukan pada domain waktu (*time domain*) maupun domain frekuensi (*frequency domain*). Domain waktu merepresentasikan perubahan amplitudo terhadap waktu, sedangkan domain frekuensi menggambarkan distribusi energi sinyal terhadap frekuensi tertentu. Analisis domain frekuensi sangat penting dalam pengolahan ucapan karena

karakteristik suara dan pola emosional lebih mudah diamati melalui representasi spektral dibandingkan hanya menggunakan amplitudo pada domain waktu (Alías dkk., 2016).

Transformasi sinyal dari domain waktu ke domain frekuensi umumnya dilakukan menggunakan *Fourier Transform*. Transformasi ini mengubah sinyal menjadi komponen-komponen frekuensi sehingga memudahkan proses analisis spektral dan ekstraksi fitur akustik. Bentuk matematis *Discrete Fourier Transform* (DFT) dinyatakan pada Persamaan (2.3) (Ambaye, 2020):

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi kn/N} \tag{2.3}$$

Keterangan:
$X(k)$ adalah representasi sinyal pada domain frekuensi,
$x(n)$ adalah sinyal diskrit,
N adalah jumlah sampel,
k adalah indeks frekuensi.

2.2.3. Normalisasi dan Pengolahan Awal Sinyal Audio

Pengolahan sinyal audio juga melibatkan proses normalisasi amplitudo untuk menjaga kestabilan rentang nilai sinyal selama proses komputasi. Normalisasi bertujuan mengurangi perbedaan intensitas antar rekaman sehingga model dapat lebih fokus mempelajari pola akustik dibanding variasi *volume* suara (Khamaisi dkk., 2025). Penelitian ini menerapkan normalisasi amplitudo dengan membagi setiap sampel amplitudo terhadap nilai amplitudo maksimum absolut sinyal sehingga seluruh rekaman memiliki rentang amplitudo yang lebih seragam sebelum dilakukan ekstraksi fitur (Hu dkk., 2024). Normalisasi amplitudo dapat dinyatakan pada Persamaan (2.4).

$$y_n[n] = \frac{y[n]}{max(|\ y[n]\ |) + \epsilon} \tag{2.4}$$

Keterangan:
$y_n[n]$ adalah sinyal hasil normalisasi,
$y[n]$ adalah sinyal asli,
$max(|\ y[n]\ |)$ adalah amplitudo maksimum sinyal,
ϵ adalah konstanta kecil untuk mencegah pembagian dengan nol.

Sinyal audio pada pengolahan ucapan sering mengandung bagian hening (*silence*) yang tidak membawa informasi penting terhadap karakteristik ucapan. Proses penghilangan *silence* dilakukan untuk mempertahankan bagian sinyal yang relevan saja. Penghilangan *silence* dapat meningkatkan kualitas fitur, mengurangi redundansi data, serta membantu efisiensi komputasi pada proses pelatihan model *deep learning* (Lee & Kwon, 2020).

Sinyal ucapan juga umumnya mengalami pelemahan energi pada frekuensi tinggi akibat karakteristik fisiologis saluran vokal manusia. Proses *pre-emphasis* digunakan untuk mengatasi kondisi tersebut dengan memperkuat komponen frekuensi tinggi pada tahap ekstraksi fitur MFCC sehingga keseimbangan spektrum sinyal menjadi lebih baik. Teknik *pre-emphasis* banyak digunakan dalam pengolahan ucapan modern karena mampu meningkatkan representasi karakteristik fonetik dan memperjelas informasi spektral pada sinyal suara. Persamaan matematisnya ditunjukkan pada Persamaan (2.5) (Rascon, 2023).

$$y[n] = x[n] - ax[n-1] \tag{2.5}$$

Keterangan:
$y[n]$ adalah sinyal hasil *pre-emphasis*,

$x[n]$ adalah sinyal masukan,
a adalah koefisien *pre-emphasis,*
$x[n-1]$ adalah sampel sebelumnya.

2.3. Augmentasi Audio

Augmentasi data merupakan teknik peningkatan jumlah dan keragaman data pelatihan dengan cara memodifikasi data asli tanpa mengubah informasi utama yang terkandung di dalamnya. *Speech Emotion Recognition* (SER) memanfaatkan augmentasi data untuk meningkatkan kemampuan generalisasi model, mengurangi *overfitting*, serta mengatasi keterbatasan jumlah data emosional yang umumnya relatif sedikit dan tidak seimbang antar kelas emosi. Teknik augmentasi juga membantu model mengenali variasi pengucapan, kondisi lingkungan, dan karakteristik suara yang berbeda-beda sehingga performa sistem menjadi lebih *robust* terhadap data baru (Madanian dkk., 2023; Tao dkk., 2023).

Augmentasi diterapkan pada pengolahan ucapan modern dengan memodifikasi karakteristik tertentu dari sinyal tanpa menghilangkan informasi emosional utama. Beberapa teknik augmentasi yang umum digunakan meliputi penambahan *noise* (*additive noise*), perubahan kecepatan bicara (*time stretching*), perubahan tinggi nada (*pitch shifting*), pergeseran waktu (*time shifting*), dan perubahan kecepatan sinyal (*speed perturbation*). Kelima teknik tersebut mampu meningkatkan variasi pola akustik sehingga model *deep learning* dapat mempelajari representasi emosi yang lebih beragam (Truong dkk., 2023).

2.3.1. Penambahan *Noise (Additive Gaussian Noise)*

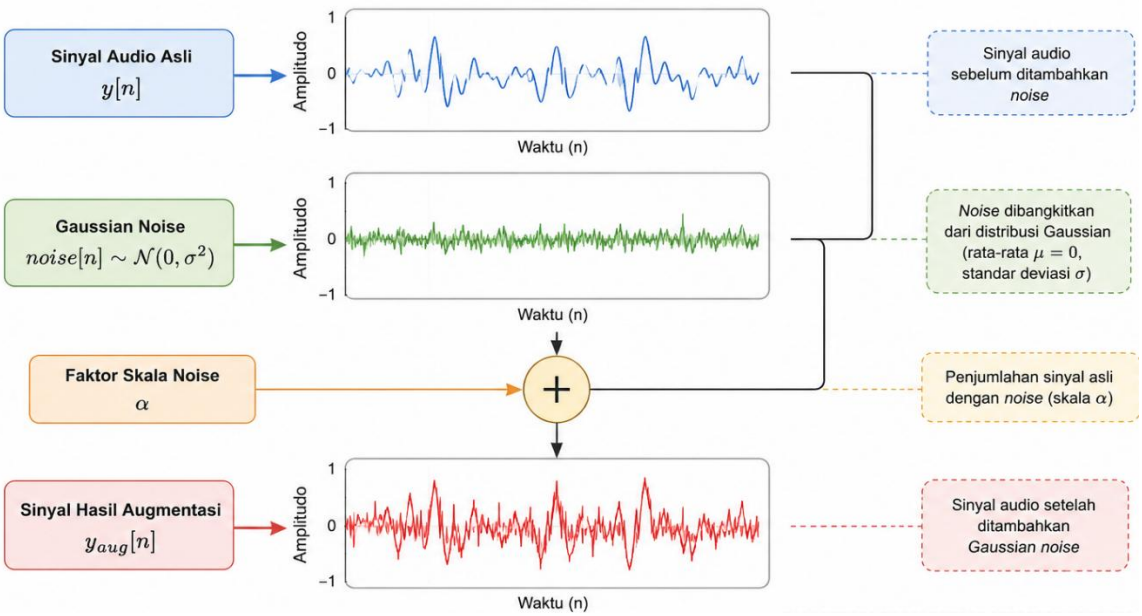
Penambahan *Additive Gaussian Noise* (AGN) merupakan teknik augmentasi data yang dilakukan dengan menambahkan sinyal *noise* acak yang mengikuti distribusi Gaussian (normal) ke dalam sinyal audio asli untuk mensimulasikan kondisi lingkungan nyata yang mengandung gangguan suara. Distribusi *Gaussian* menghasilkan nilai *noise* dengan rata-rata (μ) sebesar 0 dan standar deviasi (σ) tertentu, sehingga sebagian besar nilai *noise* memiliki amplitudo kecil dan hanya sedikit nilai yang memiliki amplitudo besar.

Penambahan *noise* dilakukan dengan mengalikan sinyal *noise* menggunakan faktor skala (α) untuk mengontrol intensitas gangguan, kemudian menjumlahkannya dengan sinyal audio asli sehingga menghasilkan sinyal hasil augmentasi. Teknik ini membantu meningkatkan variasi data latih serta membuat model lebih *robust* terhadap variasi *noise* pada proses pengujian. Augmentasi *noise* dinyatakan secara matematis pada Persamaan (2.6) (Tao dkk., 2023).

$$y_{aug}[n] = y[n] + \alpha \cdot noise[n] \tag{2.6}$$

Keterangan:
$y_{aug}[n]$ adalah sinyal hasil augmentasi,
$y[n]$ adalah sinyal asli,
$noise[n]$ adalah sinyal *noise* acak,
α adalah faktor pengontrol intensitas *noise.*

Gambar 2.2 menunjukkan ilustrasi proses penambahan *Additive Gaussian Noise* pada sinyal audio. *Gaussian noise* dibangkitkan dari distribusi normal, kemudian dikalikan dengan faktor skala (α) dan dijumlahkan dengan sinyal audio asli untuk menghasilkan sinyal hasil augmentasi (diadaptasi dari Tao dkk., 2023).



Gambar 2.2. Proses Penambahan *Additive Gaussian Noise*

2.3.2. *Time Stretching*

Time stretching merupakan teknik perubahan durasi sinyal tanpa mengubah karakteristik frekuensi atau *pitch* suara. Teknik ini digunakan untuk mensimulasikan variasi kecepatan berbicara antar pembicara. Faktor peregangan waktu dinyatakan dengan r sehingga sinyal hasil augmentasi dapat direpresentasikan sebagai Persamaan (2.7) (Truong dkk., 2023).

$$y_{stretch}(t) = y(rt) \tag{2.7}$$

Keterangan:
$y_{stretch}(t)$ adalah sinyal hasil *time stretching*,
$y[t]$ adalah sinyal asli,
r adalah faktor perubahan durasi.

2.3.3. *Pitch Shifting*

Pitch shifting merupakan teknik perubahan tinggi nada suara tanpa mengubah durasi sinyal audio. Teknik ini digunakan untuk meningkatkan variasi karakteristik vokal sehingga model dapat mengenali emosi pada berbagai tipe

suara pembicara. Pergeseran *pitch* umumnya dilakukan berdasarkan perubahan jumlah *semitone* terhadap frekuensi asli sinyal (Truong dkk., 2023).

Hubungan perubahan frekuensi terhadap jumlah *semitone* dapat dinyatakan pada Persamaan (2.8).

$$f' = f \cdot 2^{\frac{s}{12}} \tag{2.8}$$

Keterangan:
f' adalah frekuensi hasil perubahan *pitch*,
f adalah frekuensi awal,
s adalah jumlah *semitone* pergeseran.

2.3.4. *Time Shifting*

Time shifting merupakan teknik pergeseran posisi sinyal terhadap sumbu waktu tanpa mengubah isi spektral sinyal. Teknik ini bertujuan meningkatkan kemampuan model dalam mengenali pola emosi meskipun posisi temporal sinyal berubah. Secara matematis, *time shifting* dapat dituliskan pada Persamaan (2.9).

$$y_{shift}[n] = y[n - k] \tag{2.9}$$

Keterangan:
$y_{shift}[n]$ adalah sinyal hasil pergeseran,
$y[n]$ adalah sinyal asli,
k adalah jumlah pergeseran sampel.

2.3.5. Perubahan Kecepatan Sinyal (*Speed Perturbation)*

Speed perturbation merupakan teknik augmentasi yang mengubah kecepatan sinyal audio sehingga durasi dan karakteristik temporal sinyal ikut berubah. Teknik ini banyak digunakan dalam sistem pengenalan ucapan modern karena mampu meningkatkan variasi data pelatihan dan memperbaiki generalisasi model *deep learning* (Kshirsagar dkk., 2020).

2.4. Fitur Akustik

Ekstraksi fitur akustik merupakan proses mengubah sinyal suara mentah menjadi representasi numerik yang mampu menggambarkan karakteristik penting dari ucapan. Ekstraksi fitur pada sistem SER bertujuan memperoleh informasi akustik yang berkaitan dengan pola emosi pembicara, seperti perubahan frekuensi, energi, spektrum, dan dinamika temporal suara. Representasi fitur yang baik akan membantu model *deep learning* mengenali pola emosi secara lebih efektif dibandingkan menggunakan sinyal mentah secara langsung (Latif dkk., 2021).

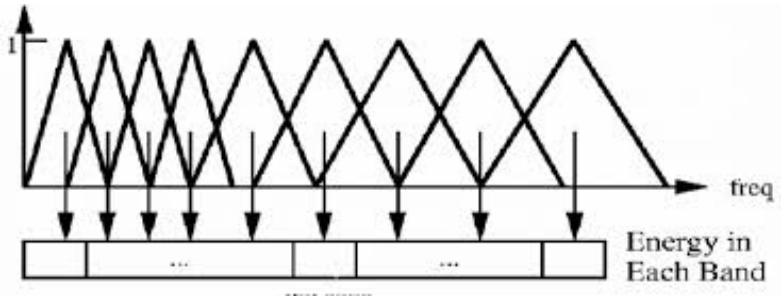
Penelitian ini menggunakan pendekatan *early fusion* dengan menggabungkan beberapa fitur akustik, yaitu MFCC, *Hilbert Multispectrum,* dan *Cochleagram Approximation*. MFCC merepresentasikan persepsi pendengaran manusia pada domain spektral, *Hilbert spectrum* menangkap karakteristik perubahan amplitudo sinyal terhadap waktu, sedangkan *Cochleagram* merepresentasikan distribusi frekuensi menyerupai mekanisme koklea pada sistem pendengaran manusia. Penggabungan beberapa fitur akustik diketahui mampu meningkatkan representasi emosi sinyal suara karena setiap fitur menyimpan informasi spektral dan temporal yang saling melengkapi (Sun, 2022).

2.4.1. *Mel Frequency Cepstral Coefficients* (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) merupakan salah satu metode ekstraksi fitur yang paling banyak digunakan dalam pengolahan ucapan karena mampu merepresentasikan karakteristik spektral suara berdasarkan persepsi pendengaran manusia. MFCC bekerja dengan mengubah sinyal suara dari domain waktu ke domain frekuensi menggunakan transformasi *Fourier,* kemudian memetakan frekuensi tersebut ke skala Mel yang menyerupai sensitivitas telinga

manusia terhadap perubahan frekuensi (Muda dkk., 2010).

Proses pembentukan MFCC diawali dengan analisis sinyal ucapan pada domain frekuensi, kemudian dilewatkan ke dalam *Mel filter bank* yang direpresentasikan pada **Gambar 2.3** untuk memperoleh distribusi energi pada setiap pita frekuensi.



Gambar 2.3. *Mel Filter Bank* (Muda dkk., 2010)

Frekuensi pada skala *Mel* dinyatakan pada Persamaan (2.10) (Muda dkk., 2010).

$$Mel(f) = 2595 log_{10}\left(1+\frac{f}{700}\right) \quad (2.10)$$

Keterangan:
f adalah frekuensi dalam Hertz,
$Mel(f)$ adalah frekuensi pada skala *Mel*.

Energi pada setiap *filter* dihitung setelah proses pemetaan frekuensi untuk menghasilkan representasi *cepstral* yang menggambarkan karakteristik spektral sinyal ucapan. MFCC banyak digunakan pada sistem *Speech Emotion Recognition* karena mampu menangkap pola fonetik dan variasi spektral yang berkaitan dengan perubahan emosi pada suara manusia. MFCC juga memiliki dimensi fitur yang relatif ringkas sehingga efektif digunakan sebagai masukan pada model *deep*

learning (Latif dkk., 2021).

2.4.2. *Hilbert Multispectrum*

Hilbert Multispectrum merupakan representasi spektral sinyal yang diperoleh melalui pendekatan *Hilbert Transform* untuk menganalisis karakteristik amplitudo dan frekuensi sinyal ucapan. Representasi ini banyak digunakan pada pengolahan sinyal *non-linier* dan *non-stasioner* karena mampu menggambarkan perubahan energi sinyal terhadap waktu secara lebih adaptif dibandingkan pendekatan *Fourier* konvensional. *Hilbert spectrum* digunakan pada pengolahan ucapan untuk menangkap pola dinamis sinyal suara yang berkaitan dengan karakteristik emosional pembicara (Vazhenina & Markov, 2020).

Transformasi *Hilbert* digunakan untuk membentuk *analytic signal* dari suatu sinyal *real*. Transformasi *Hilbert* didefinisikan secara matematis pada Persamaan (2.11) (He & Dellwo, 2016):

$$\hat{x}(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d\tau \tag{2.11}$$

Keterangan:
$\hat{x}(t)$ adalah hasil transformasi *hilbert*,
$x(t)$ sinyal asli,
P adalah *cauchy principal value*

Hasil Transformasi *Hilbert* digunakan untuk membentuk sinyal analitik pada Persamaan (2.12)

$$z(t) = x(t) + j\hat{x}(t) \tag{2.12}$$

Keterangan:
$\hat{z}(t)$ adalah hasil analitik,
j bilangan imajiner,
$\hat{x}(t)$ adalah hasil Tranformasi *Hilbert.*

Amplitudo sesaat (*instantaneous envelope*) dihitung dari sinyal analitik menggunakan *magnitude* sinyal kompleks pada Persamaan (2.13) (He & Dellwo, 2016):

$$A(t) = \sqrt{x^2(t) + \hat{x}^2(t)} \tag{2.13}$$

Keterangan:
$A(t)$ adalah *amplitude* sesaat,
$x(t)$ adalah sinyal asli,
$\hat{x}(t)$ adalah hasil Tranformasi *Hilbert.*

Representasi amplitudo sesaat selanjutnya dianalisis pada domain waktu-frekuensi menggunakan transformasi spektral sehingga menghasilkan representasi multispektrum sinyal ucapan. Pendekatan ini mampu mempertahankan informasi temporal dan spektral secara bersamaan sehingga efektif digunakan dalam pengolahan ucapan berbasis emosi (Gumelar dkk., 2020).

2.4.3. *Cochleagram*

Cochleagram merupakan representasi sinyal audio yang meniru mekanisme pemrosesan suara pada koklea manusia. Representasi ini menggambarkan distribusi energi suara terhadap waktu dan frekuensi berdasarkan karakteristik pendengaran biologis sehingga banyak digunakan pada pengolahan ucapan dan pengenalan emosi suara. *Cochleagram* mampu mempertahankan informasi temporal dan harmonik secara lebih adaptif dibandingkan representasi

spektral konvensional. *Cochleagram* sering digunakan untuk meningkatkan representasi fitur emosional pada sistem *Speech Emotion Recognition* (SER) (Peng dkk., 2023).

Penelitian ini menggunakan representasi *Cochleagram* yang tidak dibangun menggunakan *Gammatone Filterbank* biologis murni sebagaimana model koklea manusia yang umum digunakan pada *auditory modeling*. Pendekatan *Cochleagram Approximation* berbasis *Constant-Q Transform* (CQT) dipilih sebagai alternatif yang lebih efisien secara komputasi (Ong dkk., 2023). *Constant-Q Transform* (CQT) dinyatakan secara matematis pada Persamaan (2.14).

$$X(k) = \sum_{n=0}^{N_k-1} x(n) w_k(n) e^{-j2\pi Qn/N_k} \tag{2.14}$$

Keterangan:
$X(k)$ adalah koefisien CQT,
$x(n)$ adalah sinyal audio,
$w_k(n)$ adalah fungsi jendela
Q adalah *quality factor.*

Proses CQT dilakukan dengan mengalikan sinyal audio terhadap fungsi jendela pada setiap bin frekuensi sebagaimana ditunjukkan pada Persamaan (2.14). Transformasi kompleks selanjutnya dihitung menggunakan nilai *quality factor* (Q) yang konstan. Hasil transformasi *Constant-Q* selanjutnya direpresentasikan dalam bentuk energi logaritmik untuk memperjelas distribusi spektral sinyal. Representasi energi logaritmik tersebut sering digunakan sebagai pendekatan *Cochleagram* modern pada sistem SER berbasis *deep learning* karena

mampu meningkatkan sensitivitas terhadap perubahan pola emosional pada sinyal ucapan (Peng dkk., 2023).

2.5. *Early Fusion*

Early fusion atau *feature-level fusion* merupakan metode penggabungan beberapa fitur pada tahap awal sebelum proses klasifikasi dilakukan. Teknik ini dilakukan dengan mengombinasikan beberapa representasi fitur menjadi satu vektor fitur terpadu sehingga seluruh informasi dapat diproses secara bersamaan oleh model pembelajaran. *Early fusion* digunakan pada bidang *Speech Emotion Recognition* (SER) untuk mengintegrasikan berbagai karakteristik akustik ucapan agar representasi emosi yang dihasilkan menjadi lebih lengkap dan informatif (Atmaja & Sasou, 2022).

Metode *early fusion* bekerja pada level fitur (*feature level*), berbeda dengan *late fusion* yang melakukan penggabungan pada level keputusan (*decision level*). *Early fusion* menggabungkan seluruh fitur sebelum dimasukkan ke model klasifikasi, sedangkan pada *late fusion* setiap fitur diproses secara terpisah kemudian hasil klasifikasinya digabungkan pada tahap akhir. Penggabungan sejak tahap awal memungkinkan model mempelajari korelasi antarfitur secara langsung (Shixin dkk., 2024).

Proses *early fusion* umumnya dilakukan menggunakan teknik konkatenasi fitur (*feature concatenation)*. Apabila terdapat n jenis fitur yang diekstraksi dari suatu sinyal ucapan, proses penggabungan fitur dapat dinyatakan sebagai Persamaan (2.15) (Shixin dkk., 2024).

$$F_{fusion} = [F_1 \oplus F_2 \oplus F_3 \oplus ... \oplus F_n] \quad (2.15)$$

Keterangan:
F_{fusion} merupakan fitur hasil penggabungan,
F_1, F_2, ..., F_n menyatakan operasi konkatenasi,
$\oplus$ menyatakan operasi konkatenasi.

Apabila setiap fitur memiliki dimensi:

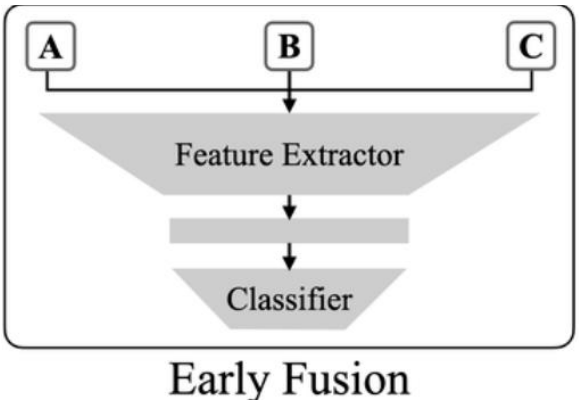
$$F_i \in \mathbb{R}^{T \times d_i} \quad (2.16)$$

Dimensi akhir *early fusion* menjadi:

$$F_{fusion} \in \mathbb{R}^{T \times \Sigma_{i=1}^{n} d_i} \quad (2.17)$$

Keterangan:
T merupakan jumlah *frame* temporal,
d_i merupakan jumlah dimensi fitur ke-i.

Alur *early fusion* secara konseptual dimulai dari proses ekstraksi beberapa fitur sinyal ucapan, kemudian seluruh fitur diselaraskan dimensinya dan digabungkan menjadi satu matriks fitur sebelum digunakan sebagai masukan model klasifikasi. **Gambar 2.4** menunjukkan ilustrasi konsep *early fusion*.



Gambar 2.4. Konsep *Early Fusion* (Orzikulova dkk., 2024)

2.6. *Padding*

Setiap rekaman audio umumnya memiliki durasi yang berbeda sehingga menghasilkan jumlah *frame* temporal yang tidak seragam setelah proses ekstraksi fitur. Kondisi ini menyebabkan setiap sampel memiliki dimensi temporal yang berbeda dan tidak dapat langsung diproses secara bersamaan dalam satu *mini-batch* pada model *deep learning.* Proses penyeragaman panjang sekuens diperlukan agar seluruh sampel memiliki dimensi yang konsisten sebelum digunakan sebagai masukan model (Iwana, 2022).

Penelitian ini menerapkan penyeragaman panjang sekuens menggunakan kombinasi *zero padding* dan *truncation*. Setiap hasil ekstraksi fitur direpresentasikan sebagai matriks pada Persamaan (2.18).

$$X \in \mathbb{R}^{T \times F} \tag{2.18}$$

T menyatakan jumlah *frame* temporal dan F menyatakan jumlah fitur pada setiap *frame.* Seluruh matriks fitur disesuaikan menjadi panjang target T_{max} sehingga memiliki dimensi yang seragam dan dapat diproses secara efisien selama pelatihan maupun pengujian model.

Panjang sekuens yang lebih pendek dari T_{max}, maka dilakukan *zero padding*, yaitu menambahkan *frame* yang seluruh elemennya bernilai nol pada bagian akhir sekuens (*post-padding*) hingga mencapai panjang target. Panjang sekuens yang melebihi T_{max}, dipotong menggunakan *post-truncation*. Pendekatan *post-padding* dipilih karena informasi pada frame asli tetap dipertahankan, sedangkan *frame* hasil *padding* hanya berfungsi sebagai pengisi untuk

menyeragamkan dimensi data (Yoon, 2020). Proses penyeragaman panjang sekuens dinyatakan sebagai Persamaan (2.19).

$$X' = \begin{cases} \begin{bmatrix} X \\ 0 \end{bmatrix}, & T < T_{max} \\ X_{1:T_{max}}, & T > T_{max} \\ X, & T = T_{max} \end{cases} \quad (2.19)$$

Keterangan:
$X \in \mathbb{R}^{T \times F}$: matriks fitur hasil ekstraksi,
T merupakan panjang sekuens awal,
0: matriks nol berukuran $(T_{max} - T) \times F$,
X': matriks fitur setelah proses penyeragaman panjang sekuens
T_{max} merupakan panjang sekuens target.

Proses tersebut menghasilkan seluruh data memiliki dimensi temporal yang sama sehingga dapat diproses secara bersamaan dalam satu *mini-batch.* Penyeragaman Panjang sekuens ini juga meningkatkan efisiensi komputasi karena seluruh sampel memiliki ukuran tensor yang konsisten selama proses pelatihan dan inferensi (Iwana, 2022).

2.7. *Convolutional Neural Network* (CNN)

Convolutional Neural Network (CNN) merupakan salah satu arsitektur *deep learning* yang digunakan untuk mengekstraksi pola dan karakteristik penting dari data melalui operasi konvolusi. CNN pertama kali dikembangkan untuk pengolahan citra, namun saat ini banyak diterapkan pada pengolahan sinyal audio dan *Speech Emotion Recognition* (SER) karena kemampuannya dalam mempelajari representasi fitur secara otomatis (LeCun dkk., 1998).

CNN digunakan pada *Speech Emotion Recognition* (SER) untuk mempelajari pola lokal dari representasi fitur akustik seperti spektrogram, MFCC,

maupun hasil *feature fusion*. CNN mampu mendeteksi pola frekuensi, perubahan energi, dan struktur spektral yang berkaitan dengan emosi ucapan. Penggunaan CNN pada SER terbukti efektif dalam meningkatkan performa klasifikasi emosi karena jaringan dapat mempelajari karakteristik emosional secara hierarkis dari data audio (Alluhaidan dkk., 2023).

CNN bekerja menggunakan operasi konvolusi antara data masukan dan kernel (*filter*) untuk menghasilkan *feature map*. Filter akan bergerak sepanjang data input untuk mendeteksi pola tertentu pada sinyal. Operasi konvolusi secara matematis dapat dituliskan sebagai Persamaan (2.20) (Krichen, 2023).

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m] \tag{2.20}$$

Keterangan:
f merupakan sinyal masukan
g merupakan kernel atau filter,
$(f * g)[n]$ merupakan hasil konvolusi.

Operasi tersebut memungkinkan CNN mempelajari representasi lokal dari data secara otomatis sehingga efektif digunakan untuk mengenali pola emosional pada sinyal ucapan. Hasil dari proses konvolusi akan membentuk *feature map* yang merepresentasikan karakteristik tertentu dari data masukan (Albawi & Mohammed, 2017).

Hasil konvolusi kemudian melewati fungsi aktivasi untuk menambahkan sifat nonlinier pada jaringan. Fungsi aktivasi yang umum digunakan pada CNN adalah *Rectified Linear Unit* (ReLU). Fungsi ReLU dapat dirumuskan pada Persamaan (2.21) (Hinton, 2010).

$$f(x) = \max(0, x) \tag{2.21}$$

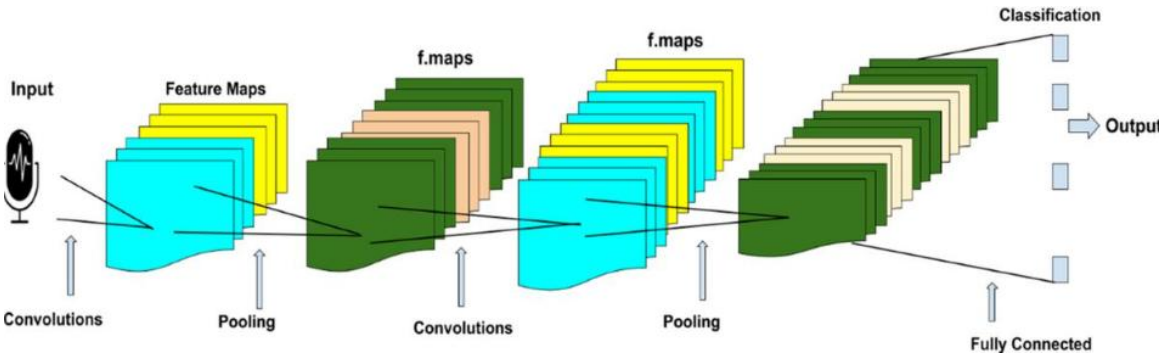
Fungsi ReLU akan mengubah nilai negatif menjadi nol dan mempertahankan nilai positif sehingga membantu mempercepat proses pelatihan model serta mengurangi masalah *vanishing gradient* (Hinton, 2010).

CNN juga menggunakan proses *pooling* untuk mengurangi dimensi *feature map* sehingga kompleksitas komputasi menjadi lebih rendah. Penelitian ini digunakan *max pooling*, yaitu proses pengambilan nilai maksimum pada suatu wilayah fitur. Operasi *max pooling* dapat dituliskan pada Persamaan (2.22).

$$y_j = max_{i \in R_j} x_i \tag{2.22}$$

Keterangan:
R_j merupakan wilayah *pooling*
x_i merupakan nilai input,
y_j merupakan hasil *pooling.*

Gambar 2.5 menunjukkan ilustrasi umum arsitektur CNN pada *Speech Emotion Recognition*.



Gambar 2.5. Ilustrasi Arsitektur *Convolutional Neural Network* (Tyagi & Szénási, 2024)

2.8. *Bidirectional Long-Short Term Memory* (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) merupakan

pengembangan dari arsitektur *Long Short-Term Memory* (LSTM) yang digunakan untuk mempelajari hubungan temporal pada data sekuensial dari dua arah, yaitu arah maju (*forward*) dan arah mundur (*backward*). BiLSTM banyak digunakan pada pengolahan sinyal audio dan *Speech Emotion Recognition* (SER) karena mampu menangkap informasi konteks masa lalu dan masa depan secara bersamaan sehingga representasi temporal yang dihasilkan menjadi lebih lengkap (Cheng dkk., 2019).

Long Short-Term Memory (LSTM) merupakan pengembangan dari *Recurrent Neural Network* (RNN) yang dirancang untuk mengatasi masalah *vanishing gradient* pada proses pelatihan data berurutan. LSTM memiliki mekanisme memori (*memory cell*) dan beberapa gerbang (*gate*) yang memungkinkan jaringan mempertahankan informasi penting dalam rentang waktu yang panjang (Hochreiter & Schmidhuber, 1997).

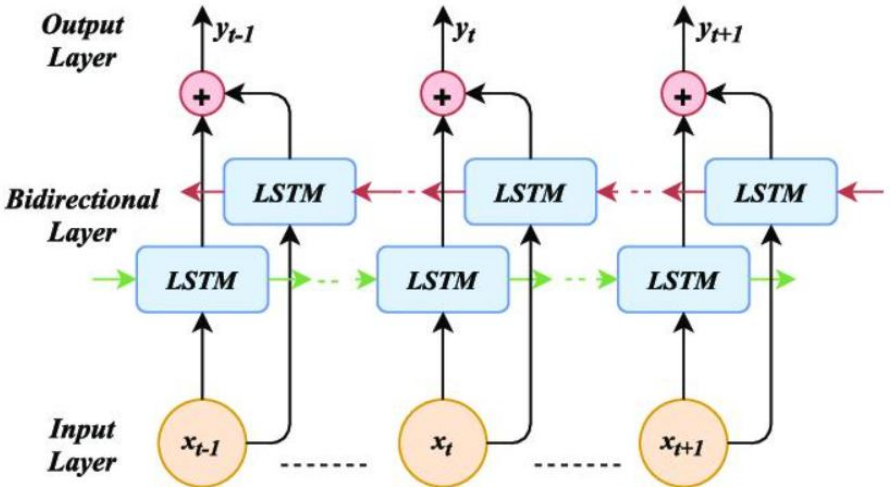
BiLSTM memproses urutan data dari dua arah secara bersamaan, berbeda dengan LSTM konvensional. Keluaran dari kedua arah dipadukan sehingga model memahami konteks temporal secara lebih menyeluruh. Representasi keluaran BiLSTM dapat dituliskan sebagai Persamaan (2.23) (Cheng dkk., 2019).

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right] \tag{2.23}$$

Keterangan:
$\overrightarrow{h_t}$ merupakan *hidden state forward*
$\overleftarrow{h_t}$ merupakan *hidden state backward*,
[;] menyatakan operasi konkatenasi.

BiLSTM digunakan pada *Speech Emotion Recognition* untuk mempelajari

dinamika temporal sinyal ucapan karena emosi pada suara dipengaruhi oleh perubahan intonasi, ritme, dan pola temporal antarframe audio. Penggunaan BiLSTM memungkinkan model memahami hubungan temporal secara lebih efektif dibandingkan metode sekuensial satu arah. Ilustrasi arsitektur BiLSTM dapat dilihat pada **Gambar 2.6**. (Olawale dkk., 2023).



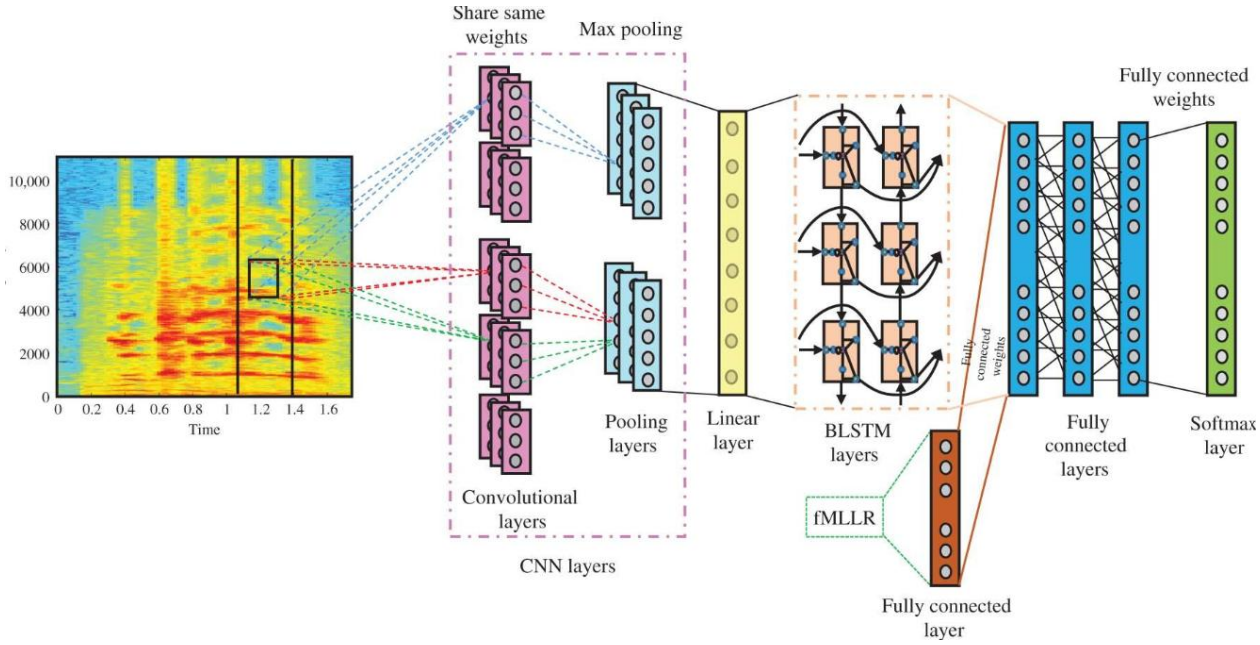
Gambar 2.6. Konsep Arsitektur BiLSTM (Olawale dkk., 2023)

2.9. CNN-BiLSTM

CNN-BiLSTM merupakan arsitektur *deep learning* hibrida yang mengombinasikan kemampuan *Convolutional Neural Network* (CNN) dalam mengekstraksi fitur spasial dengan kemampuan *Bidirectional Long Short-Term Memory* (BiLSTM) dalam mempelajari hubungan temporal pada data sekuensial. Kombinasi kedua metode tersebut banyak digunakan pada *Speech Emotion Recognition* (SER) karena sinyal ucapan memiliki karakteristik spasial dan temporal secara bersamaan (Mustaqeem & Kwon, 2020).

Gambar 2.7 menunjukkan arsitektur CNN-BiLSTM yang menggunakan CNN sebagai tahap awal ekstraksi pola lokal dari fitur akustik, seperti perubahan

energi, pola frekuensi, dan struktur spektral suara. Representasi fitur hasil CNN selanjutnya diteruskan ke BiLSTM untuk mempelajari hubungan temporal antarframe audio dari arah maju (*forward*) dan arah mundur (*backward*). Kombinasi tersebut memungkinkan model memahami dinamika emosi pada sinyal ucapan secara lebih menyeluruh.



Gambar 2.7. Konsep Arsitektur CNN-BiLSTM

Penggunaan CNN dan BiLSTM secara terintegrasi dinilai efektif karena CNN mampu mengurangi redundansi fitur dan menghasilkan representasi fitur yang lebih ringkas sebelum diproses oleh BiLSTM. BiLSTM kemudian memanfaatkan representasi tersebut untuk menangkap pola temporal jangka pendek maupun jangka panjang pada sinyal ucapan (Alluhaidan dkk., 2023).

2.10. *Temporal Attention*

Attention mechanism merupakan metode pada *deep learning* yang memungkinkan model memberikan fokus lebih besar terhadap bagian data yang

dianggap paling penting. *Temporal Attention* pada pengolahan data sekuensial memberikan fokus lebih besar karena tidak seluruh *frame* audio memiliki kontribusi yang sama terhadap emosi yang diucapkan. Mekanisme *attention* digunakan untuk memberikan bobot lebih besar pada bagian sinyal yang mengandung informasi emosional penting (Bahdanau dkk., 2015).

Temporal Attention menghitung bobot perhatian (*attention weight*) pada setiap *hidden state* yang dihasilkan oleh BiLSTM. Bobot perhatian dapat dirumuskan pada Persamaan (2.24) (Bahdanau dkk., 2015).

$$\alpha_t = \frac{exp(e_t)}{\sum_{k=1}^{T} exp(e_k)} \tag{2.24}$$

Keterangan:
α_t merupakan bobot perhatian pada waktu ke-t
e_t merupakan skor perhatian,
T merupakan panjang urutan data.

Bobot perhatian selanjutnya digunakan untuk membentuk *context vector* yang merepresentasikan informasi penting dari seluruh urutan data (Vaswani, 2017).

$$c = \sum_{t=1}^{T} \alpha_t \, h_t \tag{2.25}$$

Keterangan:
c merupakan *context vector*
h_t merupakan *hidden state* pada waktu ke-t,
a_t merupakan bobot perhatian.

Mekanisme tersebut membuat model lebih fokus pada *frame* audio yang

memiliki informasi emosional dominan dan mengurangi pengaruh *frame* yang kurang relevan. Penggunaan *Temporal Attention* terbukti mampu meningkatkan performa klasifikasi emosi karena model dapat mempelajari hubungan temporal secara lebih adaptif (Zennou, 2026).

2.11. *Softmax*

Representasi fitur *(context vector)* hasil *Temporal Attention* diteruskan ke lapisan *Fully Connected* untuk menghasilkan nilai keluaran *(logits)* bagi setiap kelas emosi. Nilai *logit* selanjutnya dipetakan menjadi distribusi probabilitas menggunakan fungsi *Softmax.* Fungsi *Softmax* merupakan fungsi aktivasi yang umum digunakan pada lapisan keluaran *(output layer)* untuk tugas klasifikasi multikelas karena mampu mengubah nilai keluaran jaringan menjadi probabilitas dengan rentang 0 hingga 1, serta jumlah keseluruhan probabilitas sama dengan 1. Fungsi *Softmax* dinyatakan pada Persamaan (2.26) (Vakalopoulou dkk., 2023).

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{c} e^{z_j}} \tag{2.26}$$

Keterangan:
$P(y_i)$ merupakan probabilitas kelas ke-i,
z_i merupakan nilai *logit* atau keluaran neuron pada kelas ke-i,
C merupakan jumlah seluruh kelas,
e merupakan konstanta eksponensial.

Persamaan tersebut menunjukkan bahwa setiap nilai logit diubah menjadi probabilitas melalui fungsi eksponensial, kemudian dinormalisasi menggunakan jumlah seluruh nilai eksponensial dari setiap kelas sehingga total probabilitas bernilai satu. Keluaran *Softmax* dapat diinterpretasikan sebagai peluang

kemunculan masing-masing kelas emosi. Kelas dengan nilai probabilitas terbesar dipilih sebagai hasil prediksi akhir model (Vakalopoulou dkk., 2023).

2.12. Evaluasi Model

Evaluasi model klasifikasi merupakan proses untuk mengukur kemampuan model dalam memetakan input ke kelas yang benar secara objektif. Evaluasi pada *Speech Emotion Recognition* tidak hanya bergantung pada akurasi, tetapi juga menggunakan metrik lain yang diturunkan dari *Confusion Matrix* untuk memberikan gambaran performa yang lebih seimbang pada data multikelas (Powers, 2011).

2.12.1. *Leave-One-Speaker-Out* (LOSO)

Leave-One-Speaker-Out (LOSO) *Cross Validation* merupakan metode evaluasi *speaker-independent* yang banyak digunakan dalam penelitian *Speech Emotion Recognition* (SER) untuk mengukur kemampuan generalisasi model terhadap pembicara yang belum pernah dilihat selama proses pelatihan (C. Lu dkk., 2024). Metode LOSO menggunakan seluruh data dari satu pembicara digunakan sebagai data pengujian, sedangkan data dari pembicara lainnya digunakan sebagai data pelatihan, kemudian proses tersebut diulang hingga setiap pembicara pernah menjadi data pengujian satu kali. Evaluasi *speaker-independent* seperti LOSO dinilai penting karena variasi karakteristik pembicara dapat memengaruhi performa model, sehingga pengujian pada pembicara yang tidak terlibat dalam proses pelatihan mampu memberikan estimasi performa yang lebih representatif terhadap kondisi nyata (Atmaja & Sasou, 2022).

2.12.2. *Confusion Matrix*

Confusion Matrix merupakan representasi tabel yang membandingkan label aktual dan prediksi model dalam bentuk *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Matriks ini menjadi dasar utama dalam evaluasi klasifikasi karena seluruh metrik seperti *precision*, *recall*, dan *F1-score* diturunkan dari elemen-elemen tersebut, serta sangat efektif untuk menganalisis kesalahan pada setiap kelas (Bagli & Visani, 2020).

2.12.3. *Accuracy, Precision, Recall,* dan *F1-Score*

Accuracy merupakan rasio prediksi benar terhadap seluruh data, dengan Persamaan (2.27) (Saito & Rehmsmeier, 2015):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.27}$$

Precision mengukur ketepatan prediksi positif, sedangkan *recall* mengukur kemampuan model dalam menemukan seluruh data positif:

$$Precision = \frac{TP}{TP + FP'} \quad Recall = \frac{TP}{TP + FN} \tag{2.28}$$

F1-score merupakan rata-rata harmonik *precision* dan *recall*:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.29}$$

F1-Score banyak digunakan dalam klasifikasi karena lebih stabil dibandingkan *accuracy*, terutama pada dataset tidak seimbang (Saito & Rehmsmeier, 2015).

Metrik *accuracy, precision, recall*, dan *F1-score* pada Persamaan (2.27)–(2.29) pada dasarnya dirancang untuk klasifikasi biner. Penelitian ini melakukan klasifikasi multi-kelas pada lima kategori emosi serta menggunakan skema evaluasi *Leave-One-Speaker-Out* (LOSO). Pendekatan *macro averaging* digunakan dengan menghitung nilai metrik untuk setiap kelas secara terpisah kemudian mengambil rata-ratanya sehingga setiap kelas memperoleh bobot yang sama tanpa dipengaruhi jumlah sampel pada masing-masing kelas (Powers, 2011).

Macro Precision dan *Recall* dihitung sebagai rata-rata nilai *precision* seluruh kelas:

$$Macro\ Precision = \frac{1}{C}\sum_{i=1}^{C} Precision_i \quad Macro\ Recall = \frac{1}{C}\sum_{i=1}^{C} Recall_i \qquad (2.30)$$

Macro F1-score diperoleh dari rata-rata nilai F1-score pada seluruh kelas:

$$Macro\ F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i \qquad (2.31)$$

2.12.4. *Classification Report*

Classification Report merupakan ringkasan evaluasi yang mencakup *precision*, *recall*, dan *F1-score* untuk setiap kelas serta rata-rata *macro* dan *weighted*. *Macro average* menghitung rata-rata tanpa mempertimbangkan jumlah sampel tiap kelas, sedangkan *weighted average* mempertimbangkan proporsi data. Pendekatan ini penting pada dataset tidak seimbang karena memberikan evaluasi yang lebih adil terhadap performa model di setiap kelas (Pedregosa dkk., 2011).

2.13. Penelitian Terkait

Penelitian terkait disajikan dalam tabel *State of The Art* (SOTA) untuk memberikan perbandingan dan gambaran terkait dengan penelitian yang telah dilakukan sebelumnya sehingga dapat diketahui posisi penelitian ini dan dapat mengidentifikasi celah penelitian berupa permasalahan, kontribusi, hasil, dan kekurangan penelitian. **Tabel 2.1** menyajikan *State of the Art* (SOTA) penelitian terkait.

Tabel 2.1. State of The Art (SOTA)

No.	Konten	Deskripsi
1.	Judul	*Improving Indonesian Speech Emotion Classification Using MFCC and BiLSTM with Audio Augmentation* (Septiyanto dkk., 2025).
	Permasalahan	Permasalahan utama yang diangkat dalam jurnal ini adalah keterbatasan penelitian *Speech Emotion Recognition* (SER) untuk bahasa Indonesia, di mana metode konvensionl seperti *Random Forest* hanya mampu mencapai akurasi sekitar 90% dan belum optimal dalam menangkap kompleksitas temporal serta pola kontekstual *bidirectional* yang khas pada prosodi bahasa Indonesia.
	Kontribusi /Keterbaharuan	Kebaruan dalam penelitian ini adalah melakukan kombinasi ekstraksi fitur MFCC, arsitektur BiLSTM, dan teknik audio *augmentation* untuk meningkatkan *robustnes* serta akurasi klasifikasi emosi suara.
	Hasil Utama	Penelitian ini menunjukkan bahwa kombinasi optimasi parameter MFCC, penerapan lima teknik augmentasi, dan arsitektur BiLSTM mampu mencapai akurasi 93.33% pada klasifikasi emosi suara bahasa Indonesia, melampaui performa metode konvensional seperti *Random Forest* dan *Gradient Boosting.*
	Kekurangan	Penelitian ini kurang mengekplorasi variasi fitur audio yang lebih kaya serta belum memanfaatkan arsitektur *hybrid* sehingga potensi peningkatkan representasi emosi dan performa model masih belum sepenuhnya digali.

No.	Konten	Deskripsi
2.	Judul	*IndoWaveSentiment: Indonesian audio dataset for emotion classification* (Bustamin dkk., 2024).
	Permasalahan	Penelitian ini mengangkat permasalahan utama berupa minimnya ketersediaan korpus audio berbahasa Indonesia untuk pengenalan emosi suara. Penelitian SER selama ini banyak bergantung pada dataset asing seperti *EmoDB* atau *RAVDESS* sehingga kurang relevan dengan prosodi, intonasi dan konteks linguistik khas bahasa Indonesia (Bustamin dkk., 2024).
	Kontribusi /Keterbaharuan	Mengembangkan dataset *IndoWaveSentiment* yaitu dataset audio berbahasa Indonesia untuk klasifikasi emosi.
	Hasil Utama	Dataset *IndoWaveSentiment* audio berbahasa Indonesia dengan 5 kategori emosi: netral, senang, terkejut, jijik, dan kecewa.
	Kekurangan	Penelitian ini hanya terbatas pada pengembangan dataset dan tidak melakukan pengembangan model menggunakan *machine learning* atau *deep learning.*
3.	Judul	Kombinasi Fitur Multispektrum *Hilbert* dan *Cochleagram* untuk Identifikasi Emosi Wicara (Gumelar dkk., 2020).
	Permasalahan	Penelitian ini menyoroti tantangan dalm keterbatasan fitur akustik tradisional yang belum mampu menangkap karakteristik unik emosi secara konsisten. Hasil ini membuat sistem sulit membedakan emosi dengan akurasi tinggi karena pola suara yang sering tumpang tindih antar kelas emosi
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penggabungan dua fitur spektrum visual, yaitu *Hilbert spectrum* (hasil *Hilbert-Huang Transform*) dan *Cochleagram*, yang meniru mekanisme pendengaran manusia. Kombinasi ini diproses dengan CNN dan LSTM, serta dilengkapi 15 fitur akustik prosodis, sehingga menghasilkan representasi emosi yang lebih kaya dibanding pendekatan sebelumnya.
	Hasil Utama	Eksperimen menggunakan tiga dataset publik (*RAVDESS, TESS, HENLO*) menunjukkan bahwa kombinasi fitur *Hilbert* dan *Cochleagram* dengan CNN mampu mencapai akurasi 90,97%, jauh lebih tinggi dibanding LSTM yang hanya mencapai 80,62%. Hal ini menegaskan keunggulan CNN dalam mengenali pola citra spektrum emosi wicara.
	Kekurangan	Penelitian ini masih terbatas pada dataset publik berbahasa Inggris, sehingga belum teruji pada bahasa lain seperti Indonesia. Selain itu, arsitektur yang digunakan hanya CNN dan LSTM tanpa eksplorasi model *hybrid* lebih mutakhir (misalnya CNN–BiLSTM dengan *attention* atau Transformer),

No.	Konten	Deskripsi
4.	Judul	*Multi-Level Attention-Based Categorical Emotion Recognition Using Modulation-Filtered Cochleagram* (Peng dkk., 2023).
	Permasalahan	Penelitian ini mengangkat permasalahan belum optimalnya fitur akustik tradisional seperti MFCC atau *Spectogram* dalam mengenali emosi wicara secara kategorikal. Fitur-fitur tersebut kurang mampu menangkap representasi spektral temporal yang kompleks sehingga akurasi sistem SER masih terbatas.
	Kontribusi /Keterbaharuan	Penggunaan *Modulation Filtered Cochleagram* (MCG) yang meniru persepsi auditori manusia, dikombinasikan dengan multi-level *attention network (channel, spatial,* dan *temporal attention).*
	Hasil Utama	Eksperimen pada dataset IEMOCAP menunjukkan bahwa metode ini mencapai *unweighted accuracy* sebesar 71% dan F1-score 69,2%, melampaui performa *baseline* seperti MFCC, *spectrogram*, maupun model *triple-attention*. Hasil ini menegaskan efektivitas MCG dengan *multi-level attention* dalam meningkatkan akurasi pengenalan emosi kategorikal.
	Kekurangan	Cakupan evaluasi yang hanya menggunakan dataset IEMOCAP dengan empat kategori emosi, sehingga generalisasi ke bahasa lain atau kondisi akustik berbeda belum teruji.
5.	Judul	*Real-Time Speech Emotion Recognition with a* CNN-BiLSTM-*Attention Deep Learning Model* (Zennou, 2026).
	Permasalahan	Penelitian ini mengangkat permasalahan pada keterbatasan sistem SER yang hanya mengandalkan fitur spektral seperti MFCC, sehingga kurang mampu menangkap variasi energi dan dinamika temporal emosi. Selain itu, banyak arsitektur sebelumnya yang belum seimbang antara akurasi tinggi dan efisiensi komputasi untuk aplikasi *real-time*
	Kontribusi /Keterbaharuan	Pengembangan arsitektur hibrid CNN–BiLSTM–*Attention* yang ringan namun efektif, dengan strategi *multi-feature fusion* menggunakan MFCC dan RMSE.
	Hasil Utama	Model yang diusulkan mencapai akurasi 91,74% pada *RAVDESS* dan 98,10% pada dataset gabungan *RavTess*, dengan *precision*, *recall*, dan *F1-score* seimbang di atas 97%. Performa ini melampaui *baseline* seperti VGG-16, *ResNet50*, dan CNN-LSTM, sekaligus menunjukkan efisiensi parameter yang relatif kecil (±240 ribu).
	Kekurangan	Belum adanya evaluasi lintas korpus secara penuh, belum membandingkan dengan model *self-supervised*

No.	Konten	Deskripsi
6.	Judul	*Machine Learning Methods for Speech Emotion Recognition* (E & Kulkarni, 2025).
	Permasalahan	Penelitian ini mengangkat permasalahan keterbatasan sistem pengenalan emosi suara yang masih bergantung pada fitur akustik sederhana (MFCC, prosodi) dan model klasik seperti SVM atau *Random Forest,* yang sulit menangkap kompleksitas temporal serta generalisasi lintas pembicara.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penerapan *feature fusion* (MFCC, *chroma, spectral contrast*) dengan augmentasi audio (*pitch shifting, time stretching, noise injection*), serta penggunaan arsitektur CNN dan *hybrid* CNN–LSTM.
	Hasil Utama	Eksperimen pada dataset *RAVDESS* menunjukkan CNN mencapai akurasi 84,2%, sementara CNN–LSTM memperoleh hasil terbaik dengan akurasi 86,4% dan *F1-score* 0,86.
	Kekurangan	Kekurangan utama penelitian ini adalah penggunaan dataset *RAVDESS* yang berbahasa Inggris, sehingga belum teruji pada bahasa Indonesia yang memiliki prosodi dan intonasi berbeda.
7.	Judul	*Emotion Detection from Speech Using* CNN-BiLSTM *with Feature Rich Audio Inputs* (Tiwari, 2025).
	Permasalahan	Keterbatasan sistem pengenalan emosi suara yang masih mengandalkan fitur akustik tradisional seperti MFCC dan prosodi, yang kurang mampu menangkap kompleksitas temporal serta variasi emosi secara konsisten. Hal ini membuat akurasi sistem SER belum optimal, terutama pada emosi yang mirip.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penerapan *feature fusion* antara MFCC, *Mel-spectrogram*, dan prosodi, dikombinasikan dengan arsitektur *hybrid* CNN–BiLSTM serta mekanisme *attention*
	Hasil Utama	Eksperimen menunjukkan bahwa model CNN–BiLSTM dengan *feature fusion* mampu mencapai akurasi 94% dibanding *baseline ensemble machine learning*. Hasil evaluasi juga memperlihatkan peningkatan signifikan pada emosi yang sebelumnya sulit dibedakan, sehingga sistem lebih *robust* dan seimbang dalam *precision*, *recall*, serta *F1-score*.
	Kekurangan	Kekurangan utama penelitian ini adalah penggunaan dataset berbahasa Inggris (*RAVDESS*), sehingga belum teruji pada bahasa Indonesia yang memiliki prosodi dan intonasi berbeda.

No.	Konten	Deskripsi
8.	Judul	*Advancing Indonesian Audio Emotion Classification: A Comparative Study Using IndoWaveSentiment* (Majiid dkk., 2025).
	Permasalahan	Penelitian ini mengangkat masalah kurangnya penelitian *Speech Emotion Recognition* (SER) berbahasa Indonesia, di mana sebagian besar studi sebelumnya menggunakan dataset asing yang tidak mencerminkan karakteristik prosodi dan intonasi khas bahasa Indonesia.
	Kontribusi /Keterbaharuan	Penelitian ini memperkenalkan *IndoWaveSentiment*, korpus baru berisi 300 rekaman audio berbahasa Indonesia dengan lima kategori emosi (senang, terkejut, jijik, kecewa, dan marah). Selain itu, penelitian ini membandingkan enam algoritma *machine learning* (*Logistic Regression*, KNN, *Gradient Boosting*, Random Forest, *Naive Bayes*, dan SVC) untuk mengidentifikasi teknik klasifikasi dan fitur akustik terbaik bagi SER bahasa Indonesia.
	Hasil Utama	Hasil eksperimen menunjukkan bahwa model *Random Forest* memberikan akurasi tertinggi, yaitu sekitar 90%, dibandingkan algoritma lain.
	Kekurangan	Pendekatan yang digunakan masih berbasis *machine learning* konvensional tanpa eksplorasi arsitektur *deep learning* seperti CNN–BiLSTM atau *Transformer*
9.	Judul	*MERaLiON-SER: Robust Speech Emotion Recognition Model for English and SEA Languages* (Sailor, 2025).
	Permasalahan	Penelitian ini mengangkat masalah rendahnya akurasi sistem pengenalan emosi suara lintas bahasa, terutama karena keterbatasan dataset yang berfokus pada bahasa tertentu dan kurangnya model yang mampu melakukan generalisasi dengan baik.
	Kontribusi /Keterbaharuan	Pengembangan *MERaLiON-SER*, sebuah kerangka kerja berbasis *multilingual representation learning* yang memanfaatkan *self-supervised learning* dan *transfer learning*.
	Hasil Utama	Hasil eksperimen menunjukkan bahwa *MERaLiON-SER* mampu meningkatkan akurasi pengenalan emosi secara signifikan dibandingkan *baseline* monolingual, dengan performa yang konsisten pada beberapa dataset multibahasa.
	Kekurangan	Evaluasi yang masih terbatas pada dataset multibahasa tertentu, sehingga belum mencakup bahasa dengan prosodi unik seperti bahasa Indonesia. Selain itu, kompleksitas arsitektur membuat kebutuhan komputasi relatif tinggi.

No.	Konten	Deskripsi
10.	Judul	*A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism* (Lieskovská dkk., 2021).
	Permasalahan	Penelitian ini menyoroti tantangan utama dalam *speech emotion recognition* (SER), yaitu keterbatasan metode tradisional yang hanya mengandalkan fitur akustik sederhana sehingga sulit menangkap kompleksitas temporal dan prosodi emosi.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penerapan *deep learning* berbasis CNN dan LSTM dengan *feature fusion* antara MFCC, *Mel-spectrogram*, dan prosodi.
	Hasil Utama	Eksperimen menunjukkan bahwa kombinasi CNN–LSTM dengan *feature fusion* menghasilkan akurasi lebih tinggi dibandingkan baseline yang hanya menggunakan satu jenis fitur atau arsitektur tunggal.
	Kekurangan	Kekurangan penelitian ini adalah penggunaan dataset berbahasa Inggris IEMOCAP, sehingga belum teruji pada bahasa lain seperti Indonesia yang memiliki prosodi dan intonasi berbeda.
11.	Judul	*Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep* BiLSTM (Sajjad & Kwon, 2020).
	Permasalahan	Penelitian ini menyoroti keterbatasan sistem pengenalan emosi suara yang masih mengandalkan fitur akustik konvensional seperti MFCC dan prosodi, sehingga kurang mampu menangkap kompleksitas temporal serta variasi emosi secara konsisten.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penerapan arsitektur *deep learning* berbasis CNN dan LSTM dengan *feature fusion* antara MFCC, *Mel-spectrogram*, dan prosodi.
	Hasil Utama	Eksperimen menunjukkan bahwa kombinasi CNN–LSTM dengan *feature fusion* menghasilkan akurasi lebih tinggi dibandingkan baseline yang hanya menggunakan satu jenis fitur atau arsitektur tunggal.
	Kekurangan	Kekurangan utama penelitian ini adalah penggunaan dataset berbahasa Inggris (misalnya *RAVDESS* atau *EmoDB*), sehingga belum teruji pada bahasa lain seperti Indonesia yang memiliki prosodi dan intonasi berbeda.

No.	Konten	Deskripsi
12.	Judul	*Speech Emotion Recognition of Indonesian Movies by Using Convolutional Neural Network* (Budi dkk., 2025).
	Permasalahan	Penelitian ini berfokus pada permasalahan terkait keterbatasan fitur akustik yang digunakan sehingga kurang mampu menangkap kompleksitas temporal dan prosodi emosi.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penerapan *feature fusion* antara MFCC, Mel-spectrogram, dan prosodi, dikombinasikan dengan arsitektur CNN–BiLSTM.
	Hasil Utama	Eksperimen menunjukkan bahwa model CNN–BiLSTM dengan *feature fusion* menghasilkan akurasi lebih tinggi dibanding baseline CNN atau LSTM tunggal.
	Kekurangan	Arsitektur yang dipakai masih CNN–BiLSTM standar tanpa mekanisme *attention*, sehingga model kurang mampu menyoroti segmen suara yang paling emosional.
13.	Judul	*Speech Emotion Recognition with Dynamic CNN and Bi-LSTM* (Finch, 2024)
	Permasalahan	Penelitian ini mengangkat permasalahan mengenai keterbatasan CNN konvensional yang menggunakan kernel statis, sehingga kurang adaptif terhadap variasi prosodi dan konten ujaran.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah pengembangan arsitektur *Dynamic CNN* (DyCNN) yang mampu menyesuaikan kernel konvolusi secara adaptif terhadap input, dikombinasikan dengan Bi-LSTM untuk menangkap konteks temporal dua arah. Ditambah dengan mekanisme *attention*, sistem dapat menyoroti segmen emosional paling relevan, sehingga lebih efektif dalam mengekstraksi fitur emosi global.
	Hasil Utama	Eksperimen pada tiga dataset (CASIA berbahasa Mandarin, EmoDB berbahasa Jerman, dan IEMOCAP berbahasa Inggris) menunjukkan peningkatan akurasi rata-rata masing-masing 59,08%, 89,29%, dan 71,25%. Hasil ini melampaui baseline mainstream dengan margin 1–3%, menegaskan efektivitas DyCNN–BiLSTM dalam meningkatkan performa SER tanpa beban komputasi berlebih.
	Kekurangan	Kekurangan penelitian ini adalah evaluasi hanya dilakukan pada tiga dataset berbahasa asing, sehingga belum teruji pada bahasa Indonesia yang memiliki prosodi unik. Selain itu, meskipun ada peningkatan akurasi, hasil masih relatif moderat pada dataset besar seperti IEMOCAP.

No.	Konten	Deskripsi
14.	Judul	*Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network* (Alluhaidan dkk., 2023).
	Permasalahan	Keterbatasan fitur akustik tradisional seperti MFCC yang sering gagal menangkap representasi emosi secara utuh, terutama dalam kondisi kompleks dengan variasi pembicara, aksen, dan lingkungan.
	Kontribusi /Keterbaharuan	Penelitian ini memperkenalkan *hybrid features* bernama MFCCT (gabungan MFCC dan fitur domain waktu) yang kemudian diproses dengan CNN ringan.
	Hasil Utama	Eksperimen pada tiga dataset publik (Emo-DB, SAVEE, RAVDESS) menunjukkan peningkatan akurasi signifikan: 96,6% untuk Emo-DB, 92,6% untuk SAVEE, dan 91,4% untuk RAVDESS.
	Kekurangan	Kekurangan penelitian ini adalah evaluasi terbatas pada dataset berbahasa Jerman dan Inggris, sehingga belum teruji pada bahasa lain seperti Indonesia yang memiliki prosodi unik.
15.	Judul	*An Autoencoder-based Feature Level Fusion for Speech Emotion Recognition* (Shixin dkk., 2024).
	Permasalahan	Keterbatasan sistem SER yang masih bergantung pada fitur akustik tradisional dan arsitektur sederhana, sehingga sulit menangkap kompleksitas temporal emosi serta generalisasi lintas pembicara.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah pengembangan arsitektur hibrid berbasis CNN–BiLSTM dengan mekanisme *attention*, serta penerapan *feature fusion* antara MFCC, Mel-spectrogram, dan prosodi.
	Hasil Utama	Eksperimen pada dataset benchmark menunjukkan peningkatan akurasi signifikan dibanding baseline CNN atau LSTM tunggal. Model CNN–BiLSTM–*Attention* dengan *feature fusion* mampu mencapai keseimbangan lebih baik pada *precision, recall,* dan *F1-score*, serta lebih robust dalam mengenali emosi yang sebelumnya sulit dibedakan.
	Kekurangan	Dataset masih menggunakan IEMOCAP dan EmoDB yang merupakan dataset dengan skala besar sehingga model belum teruji pada *low-resource dataset.*

No.	Konten	Deskripsi
16.	Judul	Analisis Pengaruh Kombinasi Fitur Spektral terhadap Tingkat Akurasi *Speech Emotion Recognition* (Farid dkk., 2023).
	Permasalahan	Penelitian *Speech Emotion Recognition* (SER) sering menghasilkan akurasi yang berbeda-beda karena dipengaruhi oleh dataset, fitur yang digunakan, serta metode klasifikasi. Banyak penelitian menggunakan berbagai fitur spektral, tetapi belum diketahui secara jelas kombinasi fitur mana yang paling berpengaruh terhadap peningkatan akurasi. Oleh karena itu diperlukan analisis untuk mengetahui pengaruh kombinasi fitur terhadap performa model SER.
	Kontribusi /Keterbaharuan	Penelitian ini menganalisis pengaruh kombinasi fitur spektral *Low Level Descriptor* (LLD) dan *High Statistical Function* (HSF) terhadap performa sistem SER serta mengidentifikasi fitur yang paling berpengaruh terhadap peningkatan akurasi.
	Hasil Utama	Hasil penelitian menunjukkan bahwa kombinasi fitur MFCC dan *spectral contrast* menghasilkan akurasi tertinggi sebesar 82%. Selain itu ditemukan bahwa MFCC merupakan fitur paling berpengaruh, dan penggunaan lebih banyak fitur tidak selalu meningkatkan akurasi model.
	Kekurangan	Penelitian hanya menggunakan satu dataset (RAVDESS) dan satu model klasifikasi yaitu LSTM
17.	Judul	*Multimodal Transformer Augmented Fusion for Speech Emotion Recognition* (Wang dkk., 2021).
	Permasalahan	Permasalahan utama dalam penelitian ini adalah keterbatasan penggunaan satu jenis fitur akustik dalam merepresentasikan emosi secara komprehensif. Banyak pendekatan sebelumnya masih bergantung pada fitur tunggal yang tidak mampu menangkap kompleksitas sinyal emosi.
	Kontribusi /Keterbaharuan	Penelitian ini memberikan analisis mendalam mengenai penggunaan *multi-feature fusion* dalam SER serta menunjukkan pentingnya kombinasi fitur untuk meningkatkan performa.
	Hasil Utama	Penggunaan *multi-feature fusion* secara konsisten menunjukkan peningkatan performa dibandingkan *single feature.*
	Kekurangan	Bersifat survey sehingga tidak menyajikan eksperimen spesifik pada dataset tertentu, termasuk Bahasa Indonesia.

No.	Konten	Deskripsi
18.	Judul	*Deep Hybrid* CNN-LSTM *Architecture with* Mel-MFCC-Chroma *Feature Fusion and Attention Mechanism for Enhanced Egyptian Arabic Speech Emotion Recognition* (Abd, 2026).
	Permasalahan	Ketimpangan pengembangan SER yang masih berfokus pada bahasa dengan sumber daya tinggi, sementara bahasa dengan banyak variasi dialek seperti Arab memiliki dataset terbatas dan kompleksitas linguistik yang tinggi.
	Kontribusi /Keterbaharuan	Penelitian ini berkontribusi dengan mengusulkan model CNN-BiLSTM berbasis attention yang mampu menangkap fitur spasial dan temporal secara lebih efektif, serta mengintegrasikan *multi-feature fusion* dari MFCC, *Mel-spectrogram,* dan *Chroma.*
	Hasil Utama	Hasil penelitian menunjukkan bahwa model CNN-LSTM dengan *attention* dan *multi-feature fusion* mencapai akurasi sebesar 93.64% mampu memberikan performa yang tinggi dan lebih unggul dibandingkan model tunggal, terutama pada kondisi dataset terbatas.
	Kekurangan	Jumlah kategori emosi yang digunakan relatif sedikit dan belum mencakup spektrum emosi yang lebih luas.
19.	Judul	*Emotion Detection Using Machine Learning: A Study on Human Computer Interaction* (Firdaus dkk., 2025).
	Permasalahan	Penelitian ini menyoroti keterbatasan sistem pengenalan emosi suara yang masih mengandalkan fitur akustik konvensional dan evaluasi sederhana.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah penerapan arsitektur hibrid berbasis CNN–BiLSTM dengan *feature fusion* antara MFCC dan *Mel-spectrogram*, serta integrasi mekanisme *attention*.
	Hasil Utama	Eksperimen menunjukkan bahwa CNN–BiLSTM–*Attention* dengan *feature fusion* menghasilkan akurasi lebih tinggi dibanding *baseline* CNN atau LSTM tunggal.
	Kekurangan	Fitur yang digunakan terbatas pada MFCC dan *Mel-spectrogram*, sehingga belum memanfaatkan representasi akustik yang lebih kaya seperti *Hilbert transform* atau *cochleagram*. Strategi fusi juga belum mengoptimalkan *early fusion*, sehingga interaksi antar fitur sejak tahap awal belum sepenuhnya dimanfaatkan.

No.	Konten	Deskripsi
20.	Judul	*Deep Feature Representations and Fusion Strategies for Speech Emotion Recognition from Acoustic and Linguistic Modalities: A Systematic Review* (Chaves-villota dkk., 2026).
	Permasalahan	Permasalahan yang diangkat adalah kurangnya sintesis komprehensif mengenai dataset, fitur representasi, dan strategi fusi yang digunakan dalam penelitian terkini.
	Kontribusi /Keterbaharuan	Kebaruan penelitian ini adalah melakukan *systematic review* dengan metodologi PRISMA terhadap literatur 2015–2024, khusus pada SER berbasis akustik dan linguistik
	Hasil Utama	Analisis menunjukkan bahwa integrasi fitur akustik dan linguistik menghasilkan peningkatan akurasi signifikan. Misalnya, pada dataset IEMOCAP, sistem multimodal mencapai *weighted accuracy* 85,71%, sementara pada MELD hanya 63,80%. Studi ini juga menegaskan bahwa *attention-based fusion* lebih unggul dibanding fusi tradisional, karena mampu menyoroti segmen emosional yang paling relevan.
	Kekurangan	Kekurangan utama penelitian ini adalah keterbatasan cakupan dataset yang sebagian besar berbahasa Inggris, sehingga generalisasi ke bahasa lain (misalnya Indonesia) belum teruji. Selain itu, meskipun *review* ini komprehensif, ia tidak menyajikan eksperimen baru atau validasi empiris lintas korpus, sehingga kontribusinya lebih bersifat konseptual daripada praktis.

Penelitian mengenai *Speech Emotion Recognition* (SER) telah berkembang pesat melalui pemanfaatan berbagai pendekatan *deep learning*, khususnya *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), *Bidirectional Long Short-Term Memory* (BiLSTM), *attention mechanism,* serta teknik *feature fusion.* Penelitian oleh Gumelar dkk. (2020), Alluhaidan dkk. (2023), Peng dkk. (2023), Shixin dkk. (2024), Tiwari (2025), Abd (2026), dan Firdaus dkk. (2025) menunjukkan bahwa penggunaan fitur akustik yang beragam, seperti MFCC, *Hilbert Multispectrum, Cochleagram, Mel-Spectrogram, Chroma,* maupun *prosodic features,* serta kombinasi arsitektur CNN-BiLSTM dan *attention mechanism* mampu meningkatkan performa

klasifikasi emosi suara dibandingkan penggunaan fitur atau model tunggal. Berbagai penelitian juga menunjukkan bahwa teknik augmentasi audio dan fusi fitur dapat meningkatkan kualitas representasi emosi serta kemampuan generalisasi model.

Penelitian *Speech Emotion Recognition* (SER) pada konteks bahasa Indonesia masih relatif terbatas. Bustamin dkk. (2024) mengembangkan dataset *IndoWaveSentiment* sebagai korpus emosi suara berbahasa Indonesia, sedangkan Majiid dkk. (2025), Septiyanto dkk. (2025), dan Budi dkk. (2025) mengembangkan model klasifikasi emosi menggunakan pendekatan *machine learning* maupun *deep learning*. Penelitian-penelitian tersebut masih menggunakan fitur yang relatif terbatas dan belum mengintegrasikan MFCC, *Hilbert Multispectrum,* dan *Cochleagram* melalui strategi *early fusion* dalam arsitektur CNN-BiLSTM berbasis *Temporal Attention.* Evaluasi *speaker-independent* untuk mengukur kemampuan generalisasi model terhadap pembicara baru juga masih terbatas. Penelitian ini mengusulkan model CNN-BiLSTM dengan *Temporal Attention* menggunakan *early fusion* antara MFCC, *Hilbert Multispectrum,* dan *Cochleagram* serta menerapkan skema *Leave-One-Speaker-Out* (LOSO) untuk klasifikasi emosi suara bahasa Indonesia.

Tabel 2.2 menyajikan matriks penelitian yang membandingkan penelitian yang diusulkan dengan penelitian-penelitian sebelumnya pada bidang *Speech Emotion Recognition* (SER). Matriks tersebut menunjukkan perbedaan dan persamaan penelitian berdasarkan aspek dataset, arsitektur model, penggunaan *attention mechanism*, jenis fitur yang digunakan, serta teknik augmentasi. Matriks

tersebut menunjukkan bahwa sebagian besar penelitian sebelumnya masih berfokus pada penggunaan fitur tunggal, dataset berbahasa asing, atau belum mengintegrasikan *attention mechanism* dan *m*ulti-*feature fusion* secara bersamaan. Penelitian ini mengusulkan pendekatan yang lebih komprehensif dengan menggabungkan dataset bahasa Indonesia, arsitektur CNN-BiLSTM berbasis temporal *attention, early fusion* antara MFCC, *Hilbert Multispectrum,* dan *Cochleagram*, serta penerapan berbagai teknik augmentasi audio untuk meningkatkan performa klasifikasi emosi suara.

Tabel 2.2. Matriks Penelitian

Peneliti	**Judul Penelitian**	**Dataset**		**Arsitektur**			***Temporal Attention***	**Fitur**			**Augmentasi**				
		Bahasa Indonesia	**Bahasa Asing**	**CNN**	**BiLSTM/ LSTM**	**CNN-BiLSTM**		**MFCC**	***Hilbert***	***Cochleagram***	***Add Noise***	***Speed change***	***Time shift***	***Time stretch***	***Pitch Shifting***
(Septiyanto dkk., 2025)	*Improving Indonesian Speech Emotion Classification Using MFCC and BiLSTM with Audio Augmentation*	√			√			√			√	√	√	√	√
(Gumelar dkk., 2020)	Kombinasi Fitur Multispektrum *Hilbert* dan *Cochleagram* untuk Identifikasi Emosi Wicara		√	√	√				√	√					
(Peng dkk., 2023)	*Multi-Level Attention-Based Categorical Emotion Recognition Using Modulation-Filtered Cochleagram*		√				√			√					

Peneliti	Judul Penelitian	Dataset		Arsitektur			Temporal Attention	Fitur			Augmentasi				
		Bahasa Indonesia	Bahasa Asing	CNN	BiLSTM/ LSTM	CNN-BiLSTM		MFCC	Hilbert	Cochleagram	*Add Noise*	*Speed change*	*Time shift*	*Time stretch*	*Pitch Shifting*
(Zennou, 2026)	*Real-Time Speech Emotion Recognition with a CNN-BiLSTM-Attention Deep Learning Model*		√			√	√	√							
(Tiwari, 2025)	*Emotion Detection from Speech Using CNN-BiLSTM with Feature Rich Audio Inputs*		√			√		√			√			√	√
(Alluhaidan dkk., 2023)	*Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network*		√	√				√							
(Shixin dkk., 2024)	*An Autoencoder-Based Feature Level Fusion for Speech Emotion Recognition*		√			√	√	√							
(Bustamin dkk., 2024)	*IndoWaveSentiment : Indonesian Audio Dataset for Emotion Classification*	√													
(Majiid dkk., 2025)	*Advancing Indonesian Audio Emotion Classification: A Comparative Study Using IndoWaveSentiment*	√						√							
Penelitian ini		√				√	√	√	√	√	√	√	√	√	√