Jurnal Politeknik Caltex Riau
Terbit Online pada laman https://jurnal.pcr.ac.id/index.php/jkt/
| e- ISSN : 2460-5255 (Online) | p- ISSN : 2443-4159 (Print) |

Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia

Yulistia Khoirotul Aini[1], Tri Budi Santoso[2] dan Titon Dutono[3]

[1]Politeknik Elektronika Negeri Surabaya, Teknik Elektro, email: yulistiaaini@pasca.student.pens.ac.id
[2]Politeknik Elektronika Negeri Surabaya, Teknik Telekomunikasi, email: tribudi@pens.ac.id
[3] Politeknik Elektronika Negeri Surabaya, Teknik Telekomunikasi, email: titon@pens.ac.id

Abstrak

Di dalam interaksi antara manusia dan komputer diperlukan kemampuan untuk melakukan pengenalan, penafsiran, dan memberikan respons emosi yang diekspresikan dalam ucapan. Sampai saat ini penelitan speech emotion recognition (SER) yang berbasis bahasa Indonesia masih sangat sedikit. Hal ini disebabkan keterbatasan korpus data berbahasa Indonesia untuk SER. Pada penelitian ini dibuat sistem SER dengan mengambil dataset dari TV series berbahasa Indonesia. Sistem dirancang dengan kemampuan untuk melakukan proses klasifikasi emosi, yaitu empat kelas label emosi marah, senang, netral dan sedih. Untuk implementasinya digunakan metode deep learning, yang dalam hal ini dipilih metode CNN. Pada sistem ini input berupa kombinasi dari tiga fitur, yaitu MFCC, frekuensi fundamental, dan RMSE. Dari eksperimen yang telah dijalankan telah diperoleh hasil terbaik untuk sistem SER berbahasa Indonesia dengan menggunakan input MFCC + frekuensi fundamental, yang menunjukkan tingkat akurasi sebesar 85%. Sedangkan akurasi terendah ketika menggunakan fitur MFCC + RMSE yaitu 72%. Dari study awal ini diharapkan mampu memberikan gambaran bagi para peneliti di bidang SER, tentang bagaimana memilih fitur sinyal wicara sebagai input di dalam pengujian dan mempermudah untuk langkah pengembangan penelitiannya.

Kata kunci: *Speech Emotion Recognition (SER), CNN, deep learning*

Abstract

In the interaction between humans and computers, the ability to recognize, interpret, and respond to emotions expressed in speech is needed. Until now, there is very little research for speech emotion recognition (SER) based on Indonesian. This is due to the limited corpus of Indonesian data for SER. In this study, a SER system was created by taking a dataset from an Indonesian TV series. The system is designed with the ability to carry out the process of classification of emotions, namely four classes of emotional labels angry, happy, neutral and sad. For its implementation, the deep learning method is used, which in this case the CNN method is selected. In this system the input is a combination of three features, namely MFCC, fundamental frequency, and RMSE. From the experiments that have been carried out, the best results have been obtained for the Indonesian language SER system using the MFCC input + fundamental frequency, which shows an accuracy rate of 85%. Meanwhile, the lowest accuracy when using the MFCC + RMSE feature is 72%. From this initial study, it is hoped that it will be

Dokumen diterima pada 07 April, 2021
Dipublikasikan pada 02 Juni, 2021

able to provide an overview for researchers in the SER field, about how to select speech signal features as input in testing and make it easier for the steps to develop their research.

Keywords: *Speech Emotion Recognition (SER), CNN, deep learning*

1. Pendahuluan

Kondisi emosional sesesorang merupakan faktor penting dalam interaksi antar manusia, dan memengaruhi beberapa aspek di dalam komunikasi, seperti ekspresi wajah, karakteristik suara, dan konten informasi linguistik. Di dalam aplikasi yang mensyaratkan interaksi secara cerdas antara manusia dengan komputer, diperlukan kemampuan untuk melakukan pengenalan, penafsiran, dan merespons emosi yang diekspresikan. Teknologi ini telah banyak diimplementasikan pada survei kepuasan pelanggan, *computer vision*, kecerdasan buatan, dan sebagainya. Dalam hal ini komputer diperlukan untuk berinteraksi dengan pengguna yang meniru interaksi manusia-dengan-manusia untuk mendapatkan pengalaman pengguna yang lebih baik [1,2,3].

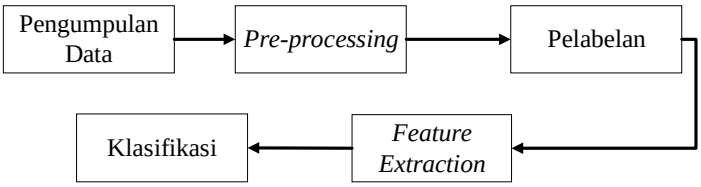
Salah satu sistem *Human–computer interaction* (HCI) dapat berupa teknologi *Speech Emotion Recognition* (SER)[4]. Beberapa contoh aplikasi penting SER di HCI adalah: sistem SER digunakan di pusat panggilan untuk mendeteksi keadaan emosional penelepon [5], sistem *smart security* di sektor perbankan, pusat panggilan cerdas dan dukungan pelanggan, medis dan aplikasi forensik, manajemen stres dan kecemasan dan mengatur pesan surat suara berdasarkan emosi [6].

Penelitian dengan topik SER telah dilakukan dengan berbagai bahasa seperti Inggris[7,8,9], Jerman[8,9], Mandarin [9], Bahasa Persia[10] dan Arab[11]. Tetapi penelitian tentang SER yang dilakukan dalam bahasa Indonesia relatif masih sedikit [12]. Salah satu penyebabnya adalah kurangnya korpus standar Bahasa Indonesia yang bisa digunakan untuk percobaan pendeteksian emosi dari suara. Selain itu, sampai saat ini SER masih menghadapi beberapa tantangan seperti tingkat akurasi yang rendah dari pengklasifikasi yang digunakan, kompleksitas komputasi yang tinggi, dan kelangkaan dalam ketersediaan natural data sets. Pengenalan emosi ucapan adalah tugas yang sulit karena beberapa alasan seperti definisi emosi yang ambigu [13] dan pemisahan yang tidak jelas antara emosi yang berbeda.

Untuk itu, pada penelitian ini telah dilakukan perancangan dan implementasi algoritma untuk pengenalan emosi pada sinyal wicara yang diambil dari TV series berbahasa Indonesia. Sistem ini disusun dengan membandingkan beberapa fitur dari sinyal suara seperti *Mel Frequency Cepstral Coeffisients* (MFCC), frekuensi fundamental dan *Root Mean Squere Energy* (RMSE). Dengan analisa beberapa fitur ini diharapkan mampu merepresentasikan emosi yang berbeda dari setiap ucapan yang memiliki ekpresi berbeda. Selanjutnya pada proses klasifikasi digunakan *Convolutional Neural Network* (CNN). Kinerja dan efisiensi komputasi dari sistem ini dianalisis menggunakan ucapan emosional bahasa Indonesia, yaitu netral, sedih, senang dan marah.

2. Metodologi

Proses deteksi emosi pada sinyal wicara dilakukan melalui beberapa tahapan proses, yang secara sederhana dapat disajikan melalui diagram blok berikut pada Gambar 1. Pada dasarnya dapat dibagi menjadi tiga proses utama. Yang pertama adalah pengumpulan data, pengolahan awal (*pre-processing*) dan pelabelan, yang kedua adalah proses ekstraksi fitur, dan yang ketiga adalah proses klasifikasi.



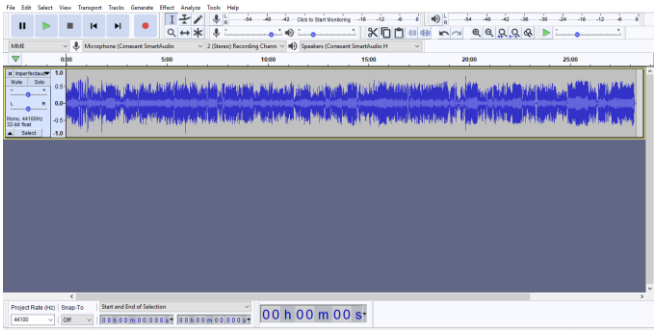
Gambar 1. Tahap metodologi penelitian

2.1 Pengumpulan Data, *Pre-processing* dan Pelabelan

Data set pada penelitian ini diambil dari TV series berbahasa Indonesia berjudul "Imperfect". Hal ini dilakukan dengan pertimbangan bahwa data dari TV series memiliki pembagian dialog terstruktur dan kualitas audio yang baik. Selain itu juga untuk menghindari terjadinya perbedaan kualitas *sound recording* dan menghindari terlalu banyaknya pembicara atau aktor.

Data yang diambil harus memenuhi kriteria berikut: suara yang diambil hanya dari satu pembicara, tidak *interference* dengan aktor lainnya, dan tidak mengandung *backsound* atau musik yang terlalu keras sehingga menutupi suara pembicara.

Pada proses *pre-processing* data yang semula berupa audio stereo (*two channels*) dirubah menjadi mono (*one channel*). Konversi dari stereo ke mono dilakukan menggunakan aplikasi "*Audacity*". Perubahan data audio dari stereo menjadi mono bertujuan untuk menyederhanakan guna mempercepat komputasi pada proses selanjutnya.



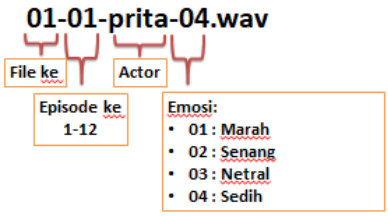
Gambar 2. Konversi stereo menjadi mono

Sample rate menunjukkan berapa banyak sampel, atau pengukuran suara yang diambil setiap detik. Semakin banyak sampel yang diambil, semakin detail gambaran naik turunnya gelombang dan semakin tinggi kualitas audionya. Selain itu, bentuk gelombang suara ditangkap dengan lebih akurat. Pada penelitian ini *sample rate* yang digunakan adalah sebesar 44100 Hz, hal ini dipilih untuk mengantisipasi munculnya suara dengan frekuensi sekitar 20 KHz. Berdasarkan hukum *Nyquist*, menyatakan bahwa frekuensi sampling minimum adalah dua kali dari frekuensi tertinggi dari suatu sinyal,untuk menghindari adanya aliasing.

Bit Depth merupakan jumlah bit informasi di setiap sampel, dan secara langsung sesuai dengan resolusi setiap sampel. Pada penelitian ini *Bit depth* yang digunakan yaitu 32 bit, sehingga memiliki jumlah jangkauan 2 pangkat 32. Semakin tinggi nilai jangkauan semakin baik kualitas audio. Tetapi ukuran file yang diperlukan juga semakin besar.

Dari file audio yang telah dikonversi menjadi mono kemudian dipotong tiap kalimat, disimpan dengan ekstensi .wav dan masing-masing diberi label berdasarkan emosi dari sinyal suara.

Pemberian label pada file disesuaikan dengan penamaan RAVDESS datasets[14]. Penamaan file pada proses pelabelan ditunjukkan pada Gambar 3.



Gambar 3. Pelabelan file audio

Proses pengumpulan data telah memberikan file audio sebanyak 507 data dengan rincian seperti yang ditunjukkan pada Tabel 1. Pemilahan jenis 4 emosi ini diharapkan sudah merepresentasikan kondisi emosi yang ada di dalam dialog TV.

Tabel 1. Rincian jumlah file hasil pelabelan

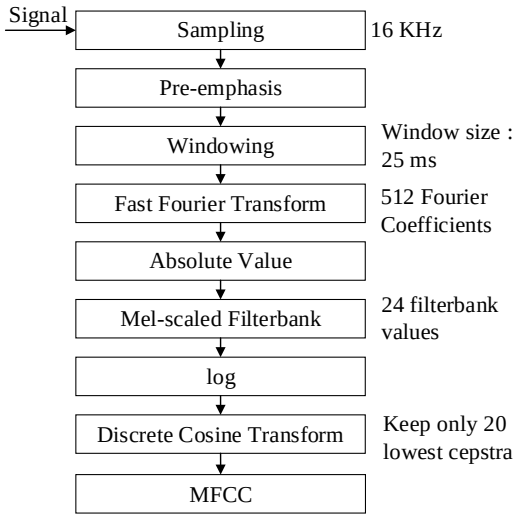
No	Emosi	Jumlah
1	Marah	154
2	Senang	104
3	Netral	141
4	Sedih	108

2.2 *Feature Extraction*

Pada penelitian ini digunakan 3 fitur suara yaitu MFCC, frekuensi fundamental, dan *Root Mean Square Energy* (RMSE). Kemudian dilakukan kombinasi terhadap ketiga fitur dengan harapan akan didapatkan akurasi yang paling baik [12].

Mel-Frequency Cepstrum Coefficient (MFCC)

MFCC memiliki resolusi frekuensi yang baik pada frekuensi rendah (< 1000 Hz), yang mana ciri fitur sinyal wicara lebih dominan di frekuensi tersebut.



Gambar 4. Proses pengolahan sinyal suara hingga didapatkan MFCC

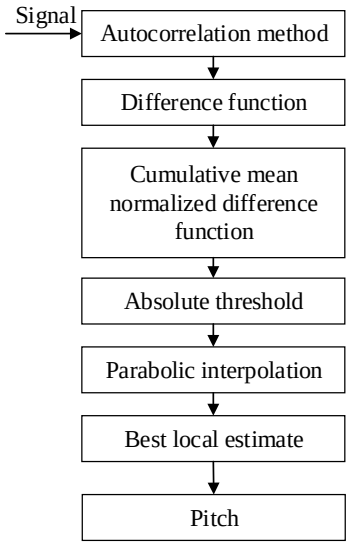
MFCC juga memiliki ketahanan terhadap kebisingan dari lingkungan. Cara kerjanya adalah dengan mengambil rata-rata atau nilai mean logaritmik spectrum setelah *Mel Filter Bank* dan *Frekuensi wrapping* [6]. Ketika seseorang sedang mengeluarkan emosi yang berbeda, maka

akan memiliki karakteristik saluran bicara yang berbeda pula, sehingga koefisien fitur MFCC dapat digunakan untuk mengidentifikasi emosi yang terkandung dalam ucapannya. Proses dari sinyal suara hingga didapatkan MFCC dapat dilihat pada Gambar 4.

Frekuensi fundamental

Frekuensi fundamental dari sinyal ucapan, sering dilambangkan dengan F_0, mengacu pada perkiraan frekuensi dari struktur periodik sinyal suara. Frekuensi fundamental didefinisikan sebagai jumlah rata-rata osilasi per detik dan dinyatakan dalam Hertz. F_0 biasanya tidak stasioner, tetapi terus berubah dalam kalimat. Pada penerapannya F_0 dapat digunakan dengan tujuan ekspresif untuk menandakan, misalnya, penekanan dan pertanyaan.

Frekuensi fundamental berkaitan erat dengan *pitch*, yang didefinisikan sebagai persepsi manusia tentang frekuensi fundamental. Artinya, F_0 menggambarkan fenomena fisik yang sebenarnya, sedangkan *pitch* menggambarkan bagaimana telinga dan otak manusia menafsirkan sinyal, dalam istilah periodisitas. Untuk mendapatkan frekuensi fundamental digunakan metode YIN dengan alur seperti pada Gambar 5.



Gambar 5. Proses pengolahan sinyal suara hingga didapatkan frekuensi fundamental

Root Mean Square Energy (RMSE)

Energi sinyal adalah besaran total sinyal, yaitu seberapa keras suatu sinyal suara. Karena amplitudo sinyal yang berosilasi bervariasi selama periode, biasanya tidak masuk akal untuk memperkirakan energi sesaat, tetapi hanya dirata-ratakan pada beberapa window. Dalam representasi *time-frequency*, energi dari komponen frekuensi tunggal dapat diperkirakan dari waktu ke waktu. Artinya, energi rata-rata dari komponen frekuensi dapat diambil melalui beberapa bingkai atau *window* berikutnya. Energi dalam sinyal didefinisikan pada persamaan 1.

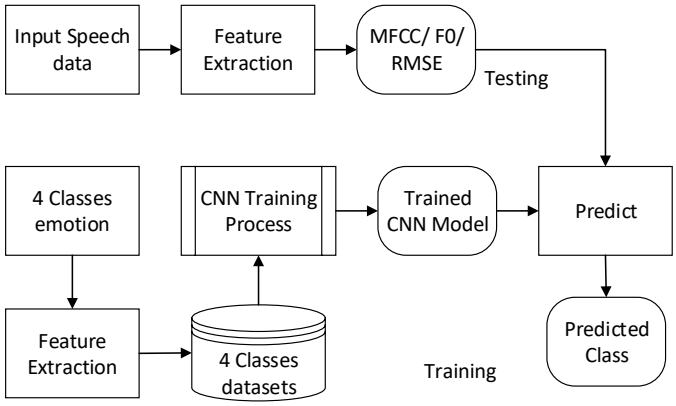
$$E(x) = \sum_n |x(n)|^2 \tag{1}$$

Root-mean-square energy (RMSE) dalam sinyal didefinisikan sebagai

$$RMSE(x) = \sqrt{\frac{1}{N}\sum_n |x(n)|^2} \tag{2}$$

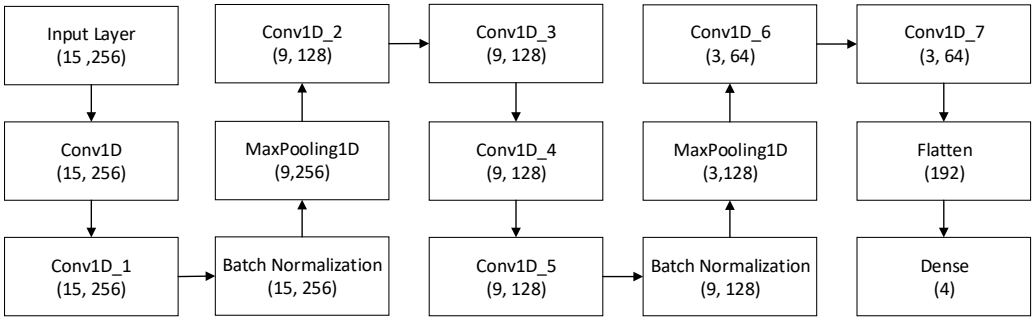
2.3 Klasifikasi

Setelah ekstraksi fitur selesai dilakukan, proses berikutnya adalah klasifikasi dengan menggunakan CNN, yang merupakan salah satu model *deep learning* untuk klasifikasi. Sistem pada penelitian ini memiliki dua buah fase yaitu fase *learning* dan fase *running* atau *testing* seperti yang disajikan pada Gambar 6.



Gambar 6. Proses training dan testing menggunakan CNN

Pada fase pembelajaran sistem menggunakan algoritma *Convolutional Neural Networks* (CNN). Pertama-tama dilakukan pengumpulan sampel emosi yang akan dikenali. Kemudian, dari semua data akan dilakukan proses fitur ekstraksi untuk mendapatkan data fitur untuk selanjutnya digunakan sebagai input pada layer CNN. Kemudian, pada layer CNN akan dilakukan proses pembelajaran. Proses pembelajaran adalah proses yang dilakukan dengan cara melakukan perubahan pada nilai parameter-parameter CNN sehingga nantinya akan didapatkan hasil nilai parameter optimum yang menyebabkan layer CNN dapat mengklasifikasi emosi dari suara yang diinputkan secara otomatis. Dan selanjutnya nilai *weight* digunakan untuk proses testing. Untuk pengujian digunakan dataset berbahasa Indonesia. Desain dari klasifikasi menggunakan *Convolutional Neural Network* ditampilkan pada Gambar 7.



Gambar 7. Desain klasifikasi Convolutional Neural Network

3. Hasil dan Pembahasan

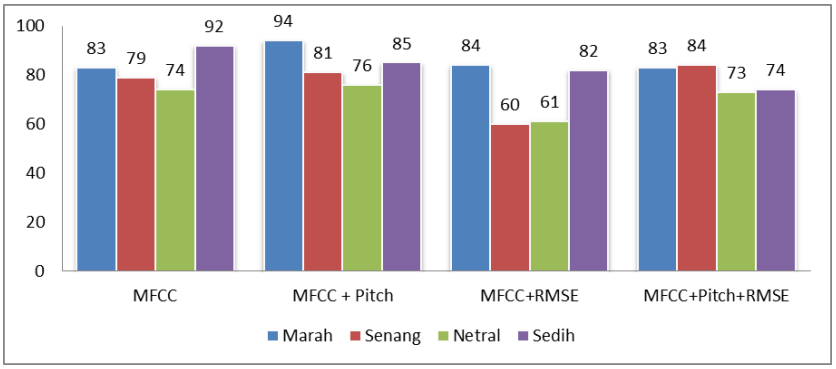
Pada bagian ini penulis menjelaskan hasil pengujian dari beberapa skenario pengujian yang telah dilakukan. Pengujian dilakukan menggunakan beberapa kombinasi fitur antara MFCC, frekuensi fundamental dan RMSE, dengan kombinasi sebagai berikut: fitur MFCC, fitur MFCC + Frekuensi Fundamental, fitur MFCC + RMSE, dan fitur MFCC + RMSE + Frekuensi Fundamental.

Metode klasifikasi yang digunakan pada penelitian ini adalah CNN dengan struktur seperti yang ditunjukkan pada Gambar 8.

Layer (type)	Output Shape
conv1d (Conv1D)	(None, 15, 256)
activation (Activation)	(None, 15, 256)
conv1d_1 (Conv1D)	(None, 15, 256)
batch_normalization (BatchNo	(None, 15, 256)
activation_1 (Activation)	(None, 15, 256)
dropout (Dropout)	(None, 15, 256)
max_pooling1d (MaxPooling1D)	(None, 9, 256)
conv1d_2 (Conv1D)	(None, 9, 128)
activation_2 (Activation)	(None, 9, 128)
conv1d_3 (Conv1D)	(None, 9, 128)
activation_3 (Activation)	(None, 9, 128)
conv1d_4 (Conv1D)	(None, 9, 128)
activation_4 (Activation)	(None, 9, 128)
conv1d_5 (Conv1D)	(None, 9, 128)
batch_normalization_1 (Batch	(None, 9, 128)
activation_5 (Activation)	(None, 9, 128)
dropout_1 (Dropout)	(None, 9, 128)
max_pooling1d_1 (MaxPooling1	(None, 3, 128)
conv1d_6 (Conv1D)	(None, 3, 64)
activation_6 (Activation)	(None, 3, 64)
conv1d_7 (Conv1D)	(None, 3, 64)
activation_7 (Activation)	(None, 3, 64)
flatten (Flatten)	(None, 192)
dense (Dense)	(None, 4)
activation_8 (Activation)	(None, 4)

Gambar 8. Desain arsitektur klasifikasi CNN

Dari total data suara sebanyak 507 file dibagi menjadi 70% untuk training dan 30% untuk testing. Untuk mengukur akurasi sistem yang diusulkan, digunakan 30% dari keseluruhan data untuk data testing. Sehingga banyak data testing untuk masing-masing label secara berurutan adalah sebanyak 50, 27, 35 dan 40 data. Pengujian dilakukan dengan epoch sebanyak 80 kali. Dari seluruh skenario yang telah dilakukan didapatkan perbandingan akurasi yang dinyatakan pada grafik pada Gambar 9.



Gambar 9. Grafik hasil pengujian sistem

Dari hasil yang didapatkan diatas dapat dianalisa bahwa kombinasi fitur paling baik untuk digunakan pada SER adalah MFCC + *pitch* dengan rata-rata akurasi sebesar 85%. Akurasi terbaik kedua adalah ketika hanya menggunakan satu fitur saja, yaitu MFCC dengan rata-rata akurasi sebesar 83%. Sementara pada pengujian dengan menggunakan fitur RMSE akurasi justru menurun baik itu MFCC + RMSE maupun MFCC + Pitch + RMSE dengan rata-rata akurasi berturut-turut sebesar 72% dan 78%. Dari sini dapat dianalisa bahwa fitur RMSE kurang baik untuk digunakan pada deteksi emosi. Hal ini dikarenakan RMSE mengambil besar energi dari suatu data *speech*. Dan diantara keempat emosi sulit untuk membedakan emosi dari keras atau tidaknya suara. Sebagai contoh ketika aktor terlalu bahagia dan bersemangat dapat mengakibatkan suara yang dihasilkan cukup keras sama seperti ketika aktor marah. Begitu pula ketika aktor sedang bersedih namun memiliki energi suara normal sehingga susah dibedakan dengan kondisi netral.

Selain itu pada TV Series “Imperfect” terdapat beberapa aktor yang memiliki aksen daerah yang berbeda-beda. Sebagai contoh ada satu tokoh bernama “Endah” yang pada TV series ini memiliki aksen Sunda, yang cenderung memiliki power suara rendah ketika berbicara meskipun pada keadaan marah. Sementara aktris lain memerankan tokoh “Prita” yang memiliki aksen Betawi, cenderung memiliki power tinggi meskipun sedang sedih. Sehingga dari hasil pengujian didapatkan fitur terbaik untuk *speech emotion recognition* adalah MFCC+Pitch.

Masing-masing pengujian memiliki *confusion matrix* seperti pada Tabel 2 hingga Tabel 5.

Tabel 2. *Confusion matrix* pengujian MFCC

Prediksi / Aktual	Marah	Senang	Netral	Sedih
Marah	37	10	3	0
Senang	0	26	1	0
Netral	2	3	23	7
Sedih	0	0	4	36

Tabel 2 menunjukkan hasil testing ketika menggunakan satu fitur untuk input, yaitu MFCC saja kesalahan prediksi terbanyak adalah pada label netral, dimana 7 diantaranya digolongkan pada kelas label sedih, 3 sebagai senang dan 2 marah. Emosi netral merupakan kelas yang paling sulit untuk dideteksi dan dibedakan dengan emosi senang dan sedih. Karena berada ditengah-tengah atau dapat dikatakan perbatasan antara masing-masing emosi.

Tabel 3. *Confusion matrix* pengujian MFCC + Pitch

Prediksi / Aktual	Marah	Senang	Netral	Sedih
Marah	49	1	2	0
Senang	3	21	2	0
Netral	0	4	30	1
Sedih	0	0	10	30

Dengan menggunakan kombinasi dua fitur yaitu MFCC dan pitch seperti yang disajikan pada Tabel 3, terjadi penurunan error pada kelas label netral ,yang sebelumnya terdapat 12 data error menjadi 5 data. Begitu pula dengan kelas label marah dan senang yang mengalami penurunan error. Kecuali pada kelas label sedih yang mengalami kenaikan error. Meskipun demikian nilai akurasi dari kombinasi input ini paling tinggi dibanding dengan kombinasi fitur yang lain.

Tabel 4. *Confusion matrix* pengujian MFCC + RMSE

Prediksi / Aktual	Marah	Senang	Netral	Sedih
Marah	37	12	1	0
Senang	1	25	1	0
Netral	0	13	17	5
Sedih	0	7	2	31

Tabel 4 menunjukkan hasil testing dengan data input MFCC dan RMSE. Pada pengujian ini terjadi peningkatan error pada kelas label marah dan netral. Terutama pada label marah yang terdeteksi menjadi senang. Hal ini dikarenakan dengan menambahkan fitur RMSE, sistem mempertimbangkan kuat lemahnya power sinyal suara. Sementara data suara marah dan senang keduanya memiliki power yang hampir sama dan menyebabkan penurunan akurasi sistem. Dan

ada beberapa data suara yang termasuk pada emosi netral, namun memiliki power suara yang cukup tinggi, sehingga terdeteksi menjadi emosi senang.

Tabel 5. *Confusion matrix* pengujian MFCC + Pitch + RMSE

Prediksi / Aktual	Marah	Senang	Netral	Sedih
Marah	39	5	5	1
Senang	2	24	0	1
Netral	3	0	31	1
Sedih	0	1	14	25

Skenario ke empat menggunakan kombinasi dari tiga fitur yaitu MFCC, Pitch dan RMSE. Seperti yang ditunjukkan pada Tabel 5, hasil prediksi emosi lebih baik dari pengujian MFCC dan RMSE, namun menurun dibanding pengujian MFCC dan Pitch saja. Hal ini menunjukkan bahwa fitur RMSE tidak merepresentasikan emosi dari suatu file suara. Penggunaan fitur RMSE justru melemahkan fitur lainnya yaitu MFCC dan pitch dalam memprediksi emosi.

Selain ketiga fitur yang digunakan pada penelitian ini, masih banyak fitur-fitur lain yang dapat digunakan sebagai ciri dari suara seperti *Linear predictive analysis* (LPC), *Linear predictive cepstral coefficients* (LPCC), *Perceptual linear predictive coefficients* (PLP) , *Relative spectra filtering of log domain coefficients* (RASTA) dan masih banyak lagi. Pada peneletian selanjutnya, diharapkan penulis dapat menganalisa pemilihan lebih banyak fitur dan pengaruh akurasi terhadap proses klasifikasi emosi.

4. Kesimpulan

Setelah melakukan perancangan, penerapan, pengujian serta analisis sistem *speech emosion recognition* dengan dataset berbahasa Indonesia, maka dapat disimpulkan bahwa:

1. Untuk mendapatkan fitur MFCC, pitch, dan RMSE semua data *speech* perlu dikonversi terlebih dahulu dari stereo menjadi mono, dengan frekuensi sampling 44100 Hz dan *bit depth* 32 bit.
2. Model CNN yang digunakan pada penelitian terdiri dari 7 layer *convolutional* 1D, 2 max pooling dan 4 *class output*. Dengan *epoch* sebanyak 80 kali, dari 507 data dibagi 70% untuk *data train* dan 30% untuk *data test.*
3. Dari pengujian dengan melakukan pengujian dengan kombinasi dari ketiga fitur didapatkan bahwa kombinasi fitur paling baik untuk digunakan pada SER adalah MFCC + pitch dengan rata-rata akurasi sebesar 85%.
4. Akurasi justru menurun ketika ditambahkan fitur RMSE, dengan akurasi terendah ketika menggunakan kombinasi fitur MFCC + RMSE yaitu sebesar 72%. Dari penelitian ini dapat disimpulkan bahwa fitur RMSE tidak sesuai untuk digunakan pada SER. Sebab RMSE mengambil besar energi dari suatu data speech, sementara energi tidak merepresentasikan emosi dari suatu sinyal wicara.

Daftar Pustaka

[1] C.M. Lee, S.S. Narayanan, "*Toward Detecting Emotions in Spoken Dialogs*", IEEE Trans, Speech Audio Process, 13(2), 293–303 , 2005.

[2] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, C. Haring, "*Activity and Emotion Recognition to Support Early Diagnosis of Psychiatric Diseases*", Second International Conference on Pervasive Computing Technologies for Healthcare, pp. 100–102, 2008.

[3] S. Yildirim, S. Narayanan, A. Potamianos, "*Detecting Emotional State of a Child in a Conversational Computer Game*". Comput. Speech Lang. 25(1), 29–44 , 2011.

[4] D. Ververidis, C. Kotropoulos, "*Emotional speech recognition: resources, features, and methods*", Speech Commun. 48 (9), 1162–1181, 2006.

[5] D.Neiberg, K. Elenius, "*Automatic Recognition of Anger in Spontaneous Speech*", INTERSPEECH 2008, Brisbane, Australia, pp. 2755–2758, 2008.

[6] Alex, S. Ben, Mary, L., & Babu, B. P. "*Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features*", Circuits, Systems, and Signal Processing, 39(11), 5681–5709, 2020.

[7] Mirsamadi, S., Barsoum, E., & Zhang, C., "*Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention Center for Robust Speech Systems"* , IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2227–2231, 2017.

[8] Mustaqeem, Sajjad, M., & Kwon, S. ,"*Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM*". IEEE Access, 8, 79861–79875, 2020.

[9] Sun, T. W., "*End-to-End Speech Emotion Recognition with Gender Information*". IEEE Access, 8, 152423–152438., 2020.

[10] Hamidi, Mina., "*Emotion Recognition from Persian Speech with Neural Network*.", International Journal of Artificial Intelligence & Applications. 3. 107-112, 2012.

[11] Hamsa, S., Shahin, I., Iraqi, Y., & Werghi, N., "*Emotion Recognition from Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier*.", IEEE Access, 8, 96994–97006, 2020.

[12] Fahmi, F., Jiwanggi, M. A., & Adriani, M. , "*Speech-Emotion Detection in an Indonesian Movie*", Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), May, 185–193, 2020.

[13] Cong, P.; Wang, C.; Ren, Z.; Wang, H.; Wang, Y.; Feng, J. "*Unsatisfied customer call detection with deep learning*", In Proceedings of the 2016 10th International Symposium on Chinese Spoken Language Processing(ISCSLP), Tianjin, China, 17–20; pp. 1–5, 2016.

[14] Livingstone, S., & Russo, F. "*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)* ". In PLoS ONE (Vol. 13), 2018.