Kombinasi Fitur Multispektrum Hilbert dan *Cochleagram* untuk Identifikasi Emosi Wicara

(*Spectrum Features Combination of Hilbert and Cochleagram for Speech Emotions Identification*)

Agustinus Bimo Gumelar[1,4], Eko Mulyanto Yuniarno[1,2], Wiwik Anggraeni[3], Indar Sugiarto[5], Andreas Agung Kristanto[6], Mauridhi Hery Purnomo[1,2]

***Abstract*—In social behavior of human interaction, human voice becomes one of the means of channeling mental states' emotional expression. Human voice is a vocal-processesed speech, arranged with word sequences, producing the speech pattern which able to channel the speakers' psychological condition. This pattern provides special characteristics that can be developed along with biometric identification process. Spectrum image visualization techniques are employed to sufficiently represent speech signal. This study aims to identify the emotion types in the human voice using a feature combination multi-spectrum Hilbert and cochleagram. The Hilbert spectrum represents the Hilbert-Huang Transformation (HHT) results for processing a non-linear, non-stationary instantaneous speech emotional signals with intrinsic mode functions. Through imitating the functions of the outer and middle ear elements, emotional speech impulses are broken down into frequencies that typically vary from the effects of their expression in the form of the cochlea continuum. The two inputs in the form of speech spectrum are processed using Convolutional Neural Networks (CNN) which best known for recognizing image data because it represents the mechanism of human retina and also Long Short-Term Memory (LSTM) method. Based on the results of this experiments using three public datasets of speech emotions, which each of them has similar eight emotional classes, this experiment obtained an accuracy of 90.97% with CNN and 80.62% with LSTM.**

[1] *Departemen Teknik Elektro, Fakultas Teknologi Elektro Dan Informatika Cerdas (ELECTICS), Institut Teknologi Sepuluh Nopember (ITS) Kampus ITS, Jl. Raya ITS, Sukolilo, Surabaya, 60111, INDONESIA (tlp/fax.: 031-5947302; email: bimo.19071@mhs.its.ac.id)*

[2] *Departemen Teknik Komputer, Fakultas Teknologi Elektro Dan Informatika Cerdas (ELECTICS), Institut Teknologi Sepuluh Nopember (ITS) Kampus ITS, Jl. Raya ITS, Sukolilo, Surabaya, 60111, INDONESIA (tlp/fax.: 031-5922936; email: hery@ee.its.ac.id; ekomulyanto@ee.its.ac.id)*

[3] *Departemen Sistem Informasi, Fakultas Teknologi Elektro Dan Informatika Cerdas (ELECTICS), Institut Teknologi Sepuluh Nopember (ITS) Kampus ITS, Jl. Raya ITS, Sukolilo, Surabaya, 60111, INDONESIA (tlp/fax.: 031-5999944; email: wiwik@is.its.ac.id)*

[4] *Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Narotama, Jl. Arif Rahman Hakim 51, Surabaya, 60117, INDONESIA (tlp/fax: 031-5946404; e-mail: bimogumelar@ieee.org)*

[5] *Program Studi Teknik Elektro, Universitas Kristen Petra, Jl. Siwalankerto No.121-131, Siwalankerto, Surabaya, 60236, INDONESIA (tlp/fax.: 031-8439040; email: indi@petra.ac.id)*

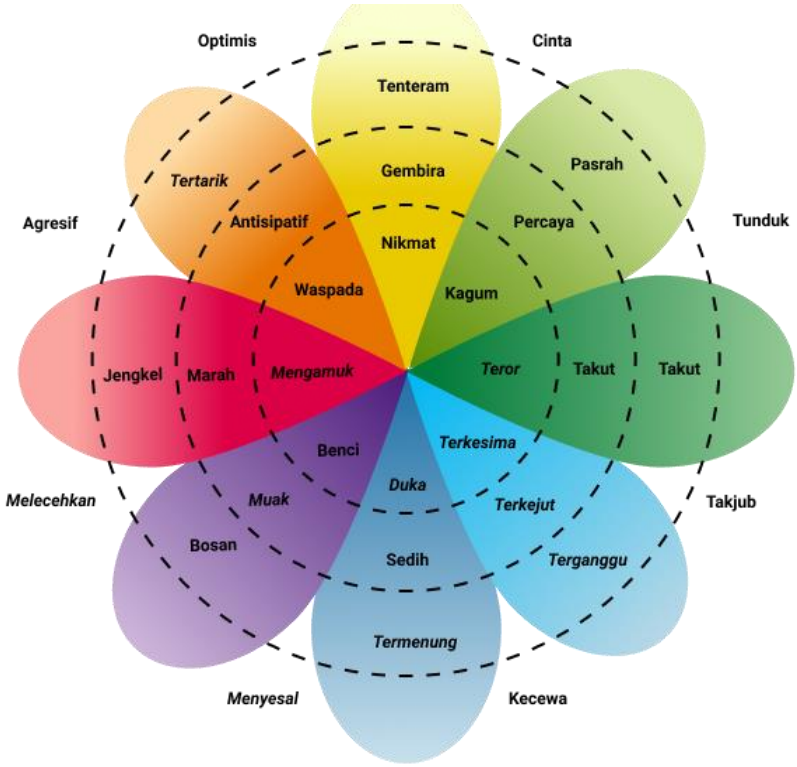
[6] *Program Studi Psikologi, Fakultas Ilmu Sosial & Ilmu Politik, Universitas Mulawarman, Jl. Tanah Grogot, Samarinda, 75411, INDONESIA (tlp/fax: 0541-743820; e-mail: andreasagungk@gmail.com)*

***Intisari*—Dalam interaksi perilaku sosial, suara manusia menjadi salah satu saluran utama pembawa atribut ekspresi emosi kondisi mentalnya. Suara manusia merupakan hasil olah vokal yang tersusun dengan disertai urutan kata demi kata, hingga menghasilkan kalimat dalam rupa pola wicara yang memiliki makna ekspresi kondisi psikologisnya. Pola tersebut memberikan karakteristik khusus untuk proses identifikasi biometrik yang menggunakan pola wicara. Teknik visualisasi berupa citra spektrum telah terbukti mampu memberikan representasi hasil olah sinyal wicara. Makalah ini mengidentifikasi jenis emosi pada wicara menggunakan kombinasi fitur multi spektrum Hilbert dan *cochleagram*. Spektrum Hilbert merepresentasikan hasil transformasi Hilbert-Huang (HHT) untuk memproses sinyal emosi wicara yang nonlinear dan nonstasioner secara instan dengan fungsi mode intrinsik. Dengan meniru cara kerja komponen telinga luar dan tengah, sinyal emosi wicara dipecah menjadi frekuensi yang berbeda secara alami dengan hasil representasinya berupa *cochleagram*. Kedua masukan berupa spektrum wicara diproses menggunakan metode *Convolutional Neural Networks* (CNN) yang dikenal terbaik dalam mengenali data citra karena merepresentasikan mekanisme kerja retina manusia, serta metode *Long Short-Term Memory* (LSTM). Berdasarkan hasil uji coba dengan tiga himpunan data (*dataset*) publik emosi wicara yang terbagi ke dalam delapan kelas emosi, diperoleh akurasi sebesar 90,97% dengan CNN dan 80,62% dengan LSTM.**

Kata Kunci*—Emosi Wicara, Kombinasi Fitur, *Convolutional Neural Networks* (CNN), *Cochleagram*, *Hilbert Spectrum*, *Deep Learning

I. PENDAHULUAN

Suara manusia (wicara) adalah salah satu alat komunikasi sosial dan afektif utama yang membawa pesan informasi bermakna dari pembicara agar dipahami dan mendapatkan respons yang berimbang oleh manusia lain yang mendengar suara tersebut di sekitarnya. Dari tahap perkembangan janin, awal kelahiran di trimester kedua, seorang bayi dapat merespons suara-suara manusia yang dipengaruhi oleh ibu sebagai ciri bahasa utama atau dasar di sekitar lingkungannya [1], [2]. Pengaruh respons suara tersebut tetap menjadi saluran utama ekspresi emosi selama perkembangan dan sepanjang hidup manusia, terlebih saat ini, mengingat interaksi sosial yang dilakukan acap kali menggunakan media telewicara, seperti perangkat telepon seluler dan media konferensi visual jarak jauh [3].



Gbr. 1 Model roda emosi versi Plutchik (*Plutchik's wheel of emotions*) [4].

Sebagian pendapat mengatakan bahwa ekspresi emosional seseorang berawal dari ekspresi wajah. Meskipun demikian, ada pula temuan yang menyatakan bahwa tidak semua ekspresi wajah dengan kondisi emosi tertentu berlaku universal. Emosi pada suatu ekspresi wajah dapat dipersepsikan lain, tergantung pada konteks budaya dan konseptual dari orang yang mempersepsikannya [5].

Rekayasa fitur suara mengalami perkembangan dari sejumlah eksperimen dengan berbagai metode *Automatic Speech Processing* (ASR) [6]–[11]. Tujuan utama peneliti menggunakan fitur suara adalah mengidentifikasi karakteristik informasi yang terkandung di dalamnya, contoh konkritnya adalah emosi. Fitur suara (akustik) pada emosi marah memiliki model spektrum yang berbeda dengan fitur suara pada emosi senang, baik secara kasat mata maupun numerik. Namun, model/bentuk ini tidaklah unik, sehingga apabila dihadapkan pada set fitur suara lainnya, tidak menghasilkan model/bentuk yang *distinctive* satu dengan lainnya. Untuk mencari informasi unik emosi dari suara manusia, "*primary key"* dari setiap emosi dalam fitur suara perlu diteliti menggunakan set fitur suara yang berbeda.

Makalah ini bertujuan melakukan proses identifikasi jenis emosi yang terkandung pada wicara dengan menggabungkan atau mengombinasikan dua fitur utama yang berbentuk citra dalam wujud spektrum, dengan penambahan fitur sekunder sejumlah lima belas fitur yang berbasis akustik dan prosodis. Kombinasi dua fitur spektrum yang diolah adalah fitur spektrum *cochleagram* dan spektrum Hilbert. Analisis penyingkapan karakteristik emosi dalam sinyal suara dapat diimprovisasi menggunakan pendekatan metode dekomposisi atas-bawah menggunakan transformasi berbasis Hilbert-Huang, atau disebut dengan *Hilbert-Huang Transform* (HHT) [12]. Kedua fitur spektrum menjadi masukan pada keseluruhan proses identifikasi jenis emosi yang dilakukan dengan langkah praproses, yaitu penentuan delapan kelas jenis emosi. Metode pembelajaran dalam algoritme jaringan saraf tiruan digunakan untuk memproses dan mempelajari pola spektrum wicara yang mengandung jenis emosi tertentu sesuai label yang telah diberikan sebelumnya. Dalam makalah ini, digunakan metode pembelajaran *Convolutional Neural Networks* (CNN) yang memiliki kinerja sangat baik dalam mengenali fitur suara dengan bentuk 2-dimensi, sedangkan pada penelitian terdahulu, *Long Short-Term Memory* (LSTM) juga banyak digunakan sebagai alat klasifikasi kelas emosi pada suara manusia yang tergolong *robust* [13]–[16].

Susunan penulisan pada makalah ini secara keseluruhan terorganisasi sebagai berikut. Bagian I menjelaskan adanya hubungan antara aspek psikologi dengan emosi pada wicara sebagai alat komunikasi sosial dan bersifat afektif. Kerangka kerja atau model psikologi emosi pada wicara yang digunakan diusulkan pada bagian II. Pada bagian selanjutnya, dilakukan proses ekstraksi fitur multi spektrum, yaitu spektrum *cochleagram* dan spektrum Hilbert yang berbasiskan HHT. Pemilihan dan penjelasan mengenai himpunan data publik yang digunakan tertuang pada bagian III. Pada bagian IV, dibahas rancangan desain dan simulasi eksperimen, juga metode evaluasi yang digunakan hingga proses analisis rangkaian eksperimen yang dilakukan dari algoritme yang diusulkan. Terakhir, bagian V menyajikan kesimpulan dari seluruh rangkaian penelitian yang dilakukan.

II. Model Psikologi Emosi Versi Plutchik

Komunikasi emosional adalah bagian penting dari interaksi sosial karena memberikan individu informasi berharga tentang keadaan orang lain serta memungkinkan orang untuk menyesuaikan perilaku dan respons dengan tepat. Ketika orang menggunakan suara untuk berkomunikasi, pendengar tidak hanya mengartikan kata-kata yang diucapkan, yaitu konten verbal, tetapi juga informasi yang terkandung, saat, dan cara kata-kata itu diucapkan, yaitu berupa konten nonverbal [4]. Sebagian besar konten nonverbal pada suara manusia dianggap untuk menyampaikan informasi tentang keadaan emosional pembicara.

Robert Plutchik sangat terkenal dengan karyanya tentang teori emosi manusia atau disebut roda emosi Plutchik [6]. Model roda emosi Plutchik ini ditunjukkan pada Gbr. 1. Plutchik menerangkan delapan emosi dasar yang masing-masing dapat dibagi tiga menurut intensitasnya, dari rendah ke tinggi, seperti terlihat pada Tabel I. Tabel II menjelaskan gabungan antar emosi dasar, sehingga membentuk emosi lanjut. Kerangka kerja roda emosi versi Plutchik ini mampu menggambarkan delapan jenis emosi dasar dan cara dari tiap emosi dapat saling berhubungan satu sama lain, termasuk pada bagian emosi tertentu yang bertentangan dan bagian emosi yang dapat dengan mudah berubah menjadi yang lain. Hal ini dapat membantu membawa kejernihan pada emosi, yang kadang-kadang dapat terasa misterius dan luar biasa.

III. Ekstraksi Fitur Multi Spektrum

Pada eksperimen ini, digunakan fitur spektrum *cochleagram*, yang selanjutnya disebut dengan *cochleagram*, dan spektrum Hilbert yang didapatkan dari hasil HHT. *Cochleagram* digunakan dengan mempertimbangkan cara kerjanya yang secara garis besar mengadopsi mekanisme pendengaran manusia [17], sedangkan HHT adalah sebuah teknik visualisasi suara sebagai hasil perbaikan dari transformasi Fourier [18].

Adapun fitur suara sekunder yang digunakan adalah *pitch*, *fundamental frequency*, dan *loudness*. Ketiga fitur suara tersebut mengalami proses perhitungan statistika *mean*, *median*, standar deviasi, *skewness*, dan *kurtosis*, sehingga berjumlah lima belas fitur suara [19].

A. Fitur Cochleagram

Penelitian sebelumnya mengembangkan beberapa bentuk visualisasi sinyal suara, seperti *cochleagram* dan *correlogram*, sebagai hasil pengembangan penelitian *spectrogram* [20]. *Spectrogram* adalah salah satu alat visualisasi sinyal suara yang kerap digunakan dalam identifikasi suara manusia, terutama untuk emosi [21]. *Spectrogram* secara bersamaan menampilkan waktu dan bentuk frekuensi dalam sinyal suara. Oleh karena sifat *spectrogram* menampilkan waktu dan frekuensi secara linear, *spectrogram* menggambarkan tingkat resolusi yang sama untuk frekuensi rendah dan frekuensi tinggi [22].

Alternatif lainnya adalah menggunakan tapis *gammatone* untuk menampilkan representasi citra dengan tingkat resolusi tinggi secara nonlinear waktu dan frekuensi dari sinyal suara, yang dikenal sebagai *cochleagram* [23]–[25]. *Cochleagram*

Tabel I
Tiga Subbagian Intensitas Delapan Emosi Dasar

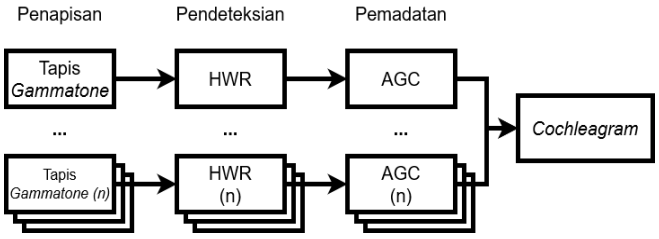
Emosi Dasar	Intensitas Emosi		
	Rendah (R)	**Sedang (S)**	**Tinggi (T)**
Kegembiraan *(Joy)*	*Serenity*	*Joy*	*Ecstasy*
Kepasrahan *(Acceptance)*	*Acceptance*	*Trust*	*Admiration*
Ketakutan *(Fear)*	*Apprehension*	*Fear*	*Terror*
Keterkejutan *(Surprise)*	*Distraction*	*Surprise*	*Amazement*
Kesedihan *(Sadness)*	*Pensiveness*	*Sadness*	*Grief*
Kemuakan *(Disgust)*	*Boredom*	*Disgust*	*Loathing*
Kemarahan *(Anger)*	*Annoyance*	*Anger*	*Rage*
Antisipasi *(Anticipation)*	*Interest*	*Anticipation*	*Vigilance*

Tabel II
Gabungan Penyusun Emosi Lanjut

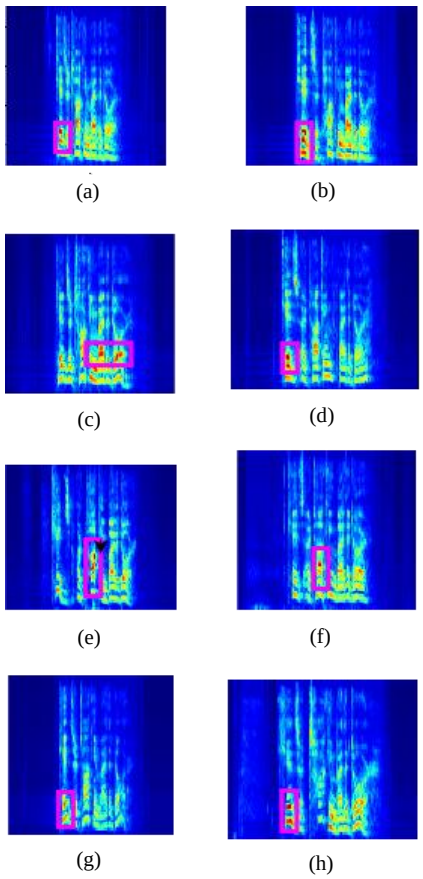
Emosi Lanjut	Emosi Dasar
Kecintaan *(Love)*	Kegembiraan *(Joy)*
	Kepercayaan (*Trust*)
Ketundukan *(Submission)*	Ketakutan *(Fear)*
	Kepercayaan (*Trust*)
Ketakjuban *(Awe)*	Keterkejutan *(Surprise)*
	Ketakutan *(Fear)*
Kekecewaan *(Dissappointment)*	Kesedihan *(Sadness)*
	Keterkejutan *(Surprise)*
Penyesalan *(Remorse)*	Kemuakan *(Disgust)*
	Kesedihan *(Sadness)*
Pelecehan *(Contempt)*	Kemarahan *(Anger)*
	Kemuakan *(Disgust)*
Keagresifan *(Aggressiveness)*	Antisipasi *(Anticipation)*
	Kemarahan *(Anger)*
Optimisme *(Optimism)*	Kegembiraan *(Joy)*
	Antisipasi *(Anticipation)*

memperkirakan informasi melalui suara yang diproses saraf pendengaran manusia dengan menyusun representasi frekuensi-waktu dari *input* sinyal suara [26]. *Cochleagram* mengubah bentuk gelombang sinyal suara *(waveform)* menjadi vektor multidimensi [17]. Tapis *gammatone* sering digunakan dengan *cochleagram*. Tapis ini merupakan tapis linear yang memodelkan properti seleksi frekuensi dari cara kerja koklea pada telinga manusia, yaitu *Basilar Membrane* (BM), dengan sel rambut di dalam telinga (*Inner Hair Cell*/IHC) yang berfungsi sebagai reseptor suara [25], [27].

Pada Gbr. 2 ditunjukkan diagram blok model koklea [17]. Tapis *gammatone* digunakan secara jamak, berfungsi sebagai penyimpanan tapis *(filterbank)*, dengan memilih frekuensi dari variasi yang diterima oleh BM dan IHC [20]. Pada implementasinya, IHC digantikan dengan *Half Wave Rectifier* (HWR) yang mendeteksi keluaran dari setiap tapis. Sifat nonlinear dari HWR menyimulasikan perubahan gerakan sebagai respons dari BM dalam koklea manusia menjadi sinyal yang merepresentasikan energi suara, dengan tetap mempertahankan informasi temporal [17], [20].



Gbr. 2 Model koklea sebagaimana diadaptasi dari Lyon [20].



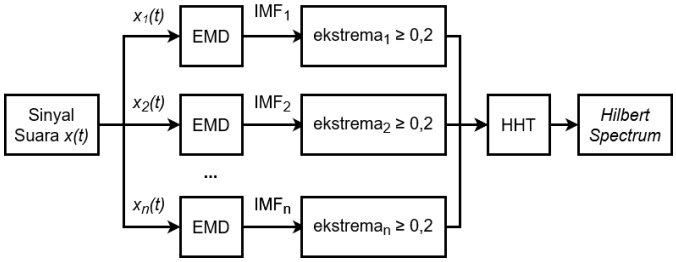
Gbr. 3 *Cochleagram* untuk emosi, (a) netral, (b) tenang, (c) senang, (d) sedih, (e) marah, (f) takut, (g) kaget, dan (h) jijik.

Pembentukan formasi *cochleagram* membutuhkan penyimpanan tapis *gammatone* yang merupakan serangkaian tapis lolos antara respon impuls, yang ditunjukkan pada (1).

$$gt(r) = Ar^{j-1}.e^{-2\pi Br}.\cos(2\pi.f_c r + \emptyset) \quad (1)$$

Nilai A merupakan amplitudo, nilai j adalah urutan tapis, nilai B adalah durasi respons impuls atau lebar pita tapis, nilai f_c adalah nilai frekuensi tengah tapis, $\emptyset$ adalah fase, dan nilai r adalah waktu. Tapis *gammatone* dilaporkan memberi perkiraan yang baik dari tapis auditori manusia untuk urutan tapis 3, 4, dan 5 [28]. Berbeda dengan *spectrogram*, tapis *gammatone* memiliki *bandwidth* tidak konstan di seluruh kanal frekuensi, yaitu *Equivalent Rectangular Bandwidth* (ERB) [29], [30]. Secara umum, ERB dari tapis *gammatone* direpresentasikan dengan (2).

$$ERB = 24{,}7(4{,}37 f_c/1000 + 1) \quad (2)$$



Gbr. 4 Diagram alir spektrum Hilbert dengan HHT.

Dalam eksperimen ini, digunakan implementasi tapis *gammatone* dari Slaney [30], [31], dengan jumlah 64 standar tapis berjarak 50 Hz hingga 8 kHz (sinyal wicara pada 16 kHz).

Tahap berikutnya yaitu tahap *Automatic Gain Control* (AGC), yang berfungsi untuk memadatkan rentang frekuensi yang luas pada suara masukan, sehingga dapat mencapai rentang frekuensi yang diterima oleh saraf pendengaran manusia [17], [20]. Gbr. 3 menunjukkan hasil citra spektrum *cochleagram* untuk setiap jenis emosi.

B. Fitur Spektrum Hilbert

HHT adalah metode analisis sinyal nonlinear dan nonstasioner, termasuk sinyal suara [32]. Dalam praktiknya, HHT memiliki dua tahap, yaitu *Empirical Mode Decomposition* (EMD) dan *Intrinsic Mode Functions* (IMF) [33]–[35]. Sebagai langkah pertama, EMD mengurai sinyal suara menjadi sejumlah IMF [36].

EMD mengidentifikasi *extrema* pada sinyal suara, lalu menghubungkan sinyal pada titik *maxima* dan *minima*, seperti terlihat pada Gbr. 4. Proses menghubungkan titik tersebut dilakukan menggunakan interpolasi *cubic spline* dan menghasilkan semacam garis bantu yang disebut dengan *envelope*. Jika sinyal suara diandaikan $x(t)$, dengan $mean_1$ bernilai satu set dari *envelope* pada batas atas (*maxima* atau *onset*) dan pada batas bawah (*minima* atau *offset*), maka IMF pertama, yaitu IMF_1, dapat dirumuskan seperti pada (3).

$$IMF_1 = x(t) - mean_1 \quad (3)$$

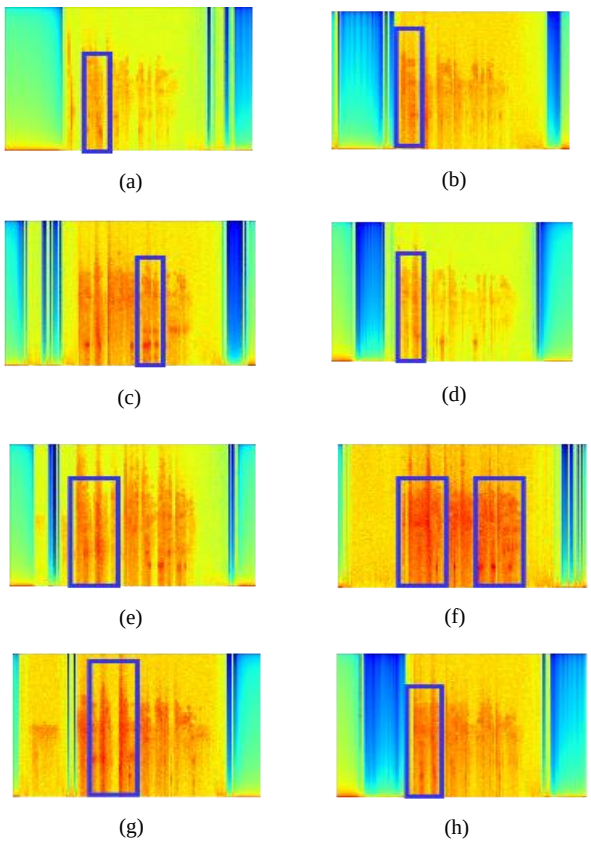
Iterasi berikutnya menggunakan IMF_1 sebagai data baru, seperti pada (4).

$$IMF_2 = IMF_1 - mean_{11} \quad (4)$$

Persamaan (4) menunjukkan $mean_{11}$, yaitu *mean* dari *envelope* atas dan bawah milik IMF_1. Pada IMF, terdapat fungsi komponen tunggal (seperti frekuensi) dengan waktu yang bervariasi [9], [32], [37]. Setiap IMF dibangikitkan menggunakan EMD dengan iterasi sampai dengan sinyal suara tidak memiliki lebih dari dua *extrema*. Pada setiap iterasi, frekuensi dihasilkan secara berurutan *(top-down)*, sehingga frekuensi tertinggi terdapat pada IMF pertama [38]. Gbr. 5 menunjukkan hasil citra spektrum Hilbert untuk setiap jenis emosi.

C. Dataset yang Digunakan

Pada eksperimen ini, digunakan beberapa data publik yang sudah teruji sebelumnya [13], [39], [40]. Data publik yang



Gbr. 5 Spektrum Hilbert untuk emosi, (a) netral, (b) tenang, (c) senang, (d) sedih, (e) marah, (f) takut, (g) jijik, dan (h) kaget.

pertama adalah data wicara dan ekspresi vokal yang mengandung emosi dalam aksen bahasa Inggris–Amerika Utara, yaitu *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) dari Departemen Psikologi di Universitas Ryerson, Toronto, Kanada [39]. Himpunan data RAVDESS berisi 4.904 suara pria dan wanita yang dikategorikan dalam delapan jenis emosi dasar manusia, yaitu netral, tenang, senang, sedih, marah, takut, jijik, dan kaget.

Himpunan data publik yang kedua menggunakan data dari *Toronto Emotional Speech Set* (TESS) dengan spesifikasi 56 orang berbicara dalam bahasa Inggris sebagai bahasa pertama dengan pengaturan ambang batas frekuensi pendengaran normal secara klinis yang dimulai dari frekuensi 250 Hz hingga 8.000 Hz [40]. Setiap peserta mendengarkan stimulus berupa rangsangan vokal yang diucapkan oleh peserta dengan usia yang lebih muda atau oleh peserta pembicara yang usianya lebih tua. Setiap peserta mendengarkan rangsangan vokal tersebut dalam jumlah yang sama dari tujuh emosi. Jumlah data berisi 2.800 data emosi suara pria dan wanita dengan kategori emosi marah, jijik, takut, bahagia, sedih, netral, dan kaget. Stimulus disajikan menggunakan perangkat bantu berupa pengeras suara yang disediakan pada bilik guna mengurangi suara pada tingkat presentasi rata-rata nilai sebesar 70 dBA. Pada fase responsi, dalam menanggapi setiap stimulus yang disajikan, para peserta menggunakan layar komputer sentuh untuk menunjukkan jenis emosi tertentu yang digambarkan pembicara.

Himpunan data publik yang ketiga adalah *Human Voice Emotion in Natural Language from On-demand Media* (HENLO). Himpunan data ini berisi data suara dari emosi dan humor berjumlah 10.000 klip suara untuk delapan emosi dasar, yaitu marah, takut, kaget, tenang, jijik, netral, sedih, dan senang. Ragam data (film, *podcast*, dan *radio streaming*) diambil dari berbagai *Media-On-Demand* (MOD). Koleksi stimulus berasal dari *script* MOD, lalu melalui proses improvisasi oleh aktor profesional. Proses *trimming* data dilakukan secara manual, kemudian dilakukan proses normalisasi data. *Rating* emosi dipresentasikan oleh psikolog profesional.

IV. Desain dan Simulasi Eksperimen

Pada eksperimen pengenalan suara ini, seluruh himpunan data publik, RAVDESS [39], TESS [40], dan HENLO, mengadopsi kelas emosi sesuai versi Robert Plutchik [4]. Himpunan data suara dalam RAVDESS, TESS, dan HENLO berjumlah 17.704 data suara dengan format *.wav*.

A. Proses Pelabelan Emosi

Sebelum melakukan ekstraksi fitur, terlebih dahulu dilakukan pelabelan himpunan data. Pelabelan pada ketiga himpunan data publik mengacu pada ketentuan dari RAVDESS sebagai berikut.

- Jenis *file* (01 = Audio-Video, 02 = video saja tanpa suara, 03 = suara saja).
- Kanal suara (01 = ucapan, 02 = lagu).
- Jenis emosi (01 = netral, 02 = tenang, 03 = bahagia, 04 = sedih, 05 = marah, 06 = takut, 07 = jijik, 08 = terkejut).
- Intensitas emosi (01 = normal, 02 = kuat), dengan catatan untuk jenis emosi netral tidak tersedia intensitas kuat.
- Ucapan (01 = *"Kids are talking by the door"*, 02 = *"Dogs are sitting by the door"*).
- Pengulangan (01 = Pengulangan Pertama, 02 = Pengulangan Kedua).
- Aktor (01 – 24, dengan nomor ganjil adalah aktor pria dan nomor genap adalah aktor wanita). Sebagai catatan, HENLO tidak memiliki pelabelan aktor.

Metode dasar pertama (*baseline*) digunakan sebagai fitur utama. Fitur yang digunakan adalah MFCC dengan vektor fitur untuk setiap *frame* adalah tiga puluh dimensi. Ketiga belas fitur MFCC menggunakan sepuluh sistem penyimpanan tapis, ditambah nilai delta dan jumlah akselerasi yang muncul. Total keseluruhan ukuran dari vektor fitur untuk sinyal adalah $30x$, dengan nilai N_t adalah jumlah *frame* dalam sinyal, yang berbeda dalam setiap kasus, tergantung pada panjang sinyal. Setelah proses normalisasi data, jumlah vektor fitur akhir 68-dimensi dibentuk dengan menggabungkan rerata dan standar deviasi untuk setiap dimensi.

B. Identifikasi Menggunakan CNN

Dalam ranah pengolahan citra dengan komputer, CNN merupakan metode *deep learning* yang terbukti dan terkenal karena kinerjanya yang sangat baik [41]–[43]. Melalui perkembangan dalam mengenali suara, CNN mulai digunakan

TABEL III
ARSITEKTUR CNN DENGAN SPEKTRUM HILBERT DAN *COCHLEAGRAM*

Layer	***Output Shape***	***Layer***	***Output Shape***
input 1 (Input Layer)	(*None*, 1, 15)	*input 2 (Input Layer)*	(*None*, 64, 192. 1)
conv1d 1 (Conv1D)	(*None*, 1, 64)	*conv2d 1 (Conv2D)*	(*None*, 62, 190, 64)
max pooling1d 1 (MaxPooling1D)	(*None*, 1, 64)	*max pooling2d1 (MaxPooling2D*	(*None*, 31, 95, 64)
batch normalization 1 (BatchNormalization)	(*None*, 1, 64)	*batch normalization 4 Batch Normalization*	(*None*, 14, 46, 32)
dropout 1 (Dropout)	(*None*, 1, 32)	*dropout 2 (Dropout)*	(*None*, 14, 46, 32)
flatten 1 (Flatten)	(*None*, 32)	*flatten 2 (Flatten)*	(*None*, 20608)
dense 1 (Dense)	(*None*, 8)	*dense 2 (Dense)*	(*None*, 8)
Layer		***Output Shape***	
concatenate 1 (Contatenate)		(*None*, 16)	
batch normalization 5 Batch Normalization		(*None*, 16)	
dropout 3 (Dropout)		(*None*, 16)	
dense 3 (Dense)		(*None*, 8)	

dalam bidang pengolahan sinyal suara dengan mengubah sinyal menjadi citra seperti *spectrogram*, *cochleagram*, dan spektrum Hilbert [43]–[45]. CNN dikenal sangat baik dalam ekstraksi fitur; memperbaiki kelemahan metode LSTM, suatu metode *deep learning* yang terkenal dalam *speech processing* (khusus sinyal suara) [44], [46]. Tabel III menunjukkan arsitektur CNN dalam eksperimen ini. Sebagai *input*, digunakan citra *input* 32x15, sebagai hasil dari masing-masing *cochleagram* dan spektrum Hilbert dari ketiga himpunan data publik.

Pada proses ini terlebih dahulu dilakukan pemilahan antara data uji dengan data latih. Pemilahan dilakukan secara acak dengan persentase jumlah data latih sampai dengan 80% (angka acak). Selanjutnya, dilakukan konversi menjadi data numerik menggunakan fitur "LabelEncoder" dari *library scikit-learn* (*sklearn*) [47]. Setelah data latih dan data uji diolah dan dipersiapkan, selanjutnya dibuat model dengan jumlah tiga puluh *layer*. Jenis model yang dibuat adalah *sequential*. Untuk melakukan proses belajar *(learning)*, terlebih dahulu model yang sudah dikonstruksi dikompilasi, lalu proses pembelajaran dimulai. Proses pembelajaran dilakukan dengan jumlah *batch* sebanyak enam belas *batch* dan jumlah *epoch* sebanyak 1.000.

C. Identifikasi Menggunakan LSTM

LSTM merupakan hasil adopsi arsitektur salah satu turunan *Neural Network* (NN), yaitu *Recurrent Neural Network* (RNN) [48], [49]. LSTM dirancang untuk mempelajari suatu urutan (*sequence*) (termasuk suara), secara kontinu berdasarkan *long-term dependency*-nya [15], [49].

TABEL IV
ARSITEKTUR LSTM DENGAN SPEKTRUM HILBERT DAN *COCHLEAGRAM*

Layer	***Output Shape***	***Layer***	***Output Shape***
input 1 (Input Layer)	(*None*, 1, 15)	*input 2 (Input Layer)*	(*None*, 64, 192)
lstm 1 (LSTM)	(*None*, 1, 500)	*lstm 3 (LSTM)*	(*None*, 64, 50)
lstm 2 (LSTM)	(*None*, 100)	*lstm 4 (LSTM)*	(*None*, 10)
dropout 1 (Dropout)	(*None*, 100)	*dropout 2 (Dropout)*	(*None*, 10)
dense 1 (Dense)	(*None*, 8)	*dense 2 (Dense)*	(*None*, 8)
Layer		***Output Shape***	
concatenate 1 (Contatenate)		(*None*, 16)	
dropout 3 (Dropout)		(*None*, 16)	
dense 3 (Dense)		(*None*, 8)	

TABEL V
HASIL EVALUASI PEMBELAJARAN (KOMPOSISI DATA 80:20)

Teknik Pembelajaran	**Akurasi Pembelajaran per Emosi (%)**							
	Nt	**Tn**	**Sn**	**Sd**	**Mr**	**Tk**	**Jk**	**Kgt**
RAVDESS								
CNN	94	83	83	89	94	94	94	94
LSTM	17	22	11	11	6	11	39	67
TESS								
CNN	95	89	87	85	96	94	91	93
LSTM	14	15	23	19	15	22	41	79
HENLO								
CNN	95	88	94	91	97	93	95	97
LSTM	19	28	37	43	29	25	44	79

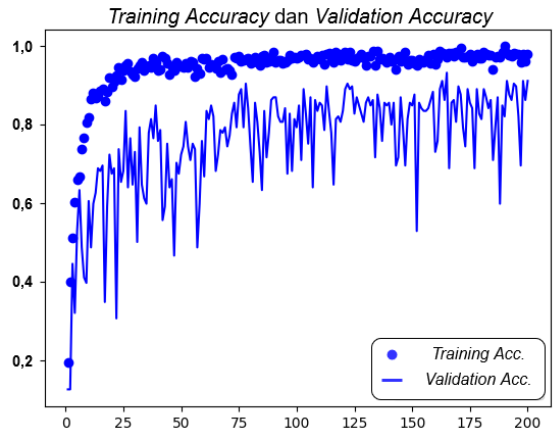
Tabel IV menunjukkan arsitektur jaringan LSTM yang digunakan dalam makalah ini untuk mengidentifikasi emosi. Pada setiap langkah waktu, terdapat hubungan langsung ke *layer* LSTM, diikuti dengan tiga *layer* LSTM yang ditumpuk secara berurutan.

D. Evaluasi

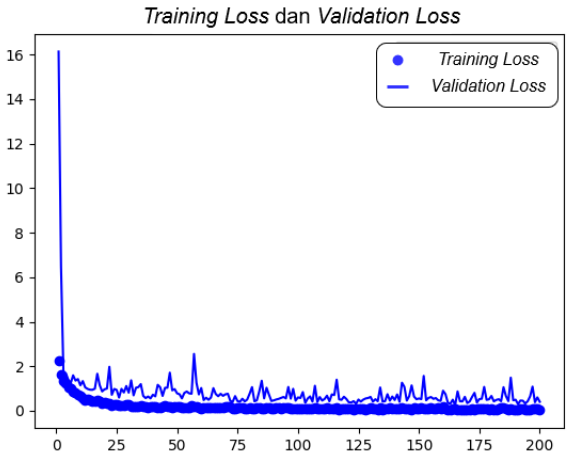
Metode *confusion matrix* merupakan representasi kinerja model klasifikasi. Matriks (tabel) menunjukkan jumlah contoh yang diklasifikasikan dengan benar dan salah, dibandingkan dengan hasil aktual (nilai target) dalam data uji. Salah satu keuntungan penggunaan *confusion matrix* sebagai alat evaluasi adalah dimungkinkannya analisis yang lebih rinci.

Dari proses pembelajaran, didapatkan hasil akhir akurasi 90,97%. Proses pembelajaran memakan waktu berkisar antara 1-2 detik per *epoch.* Hasil pembelajaran ditampilkan dalam bentuk grafik *loss*, grafik akurasi, dan *confusion matrix.*

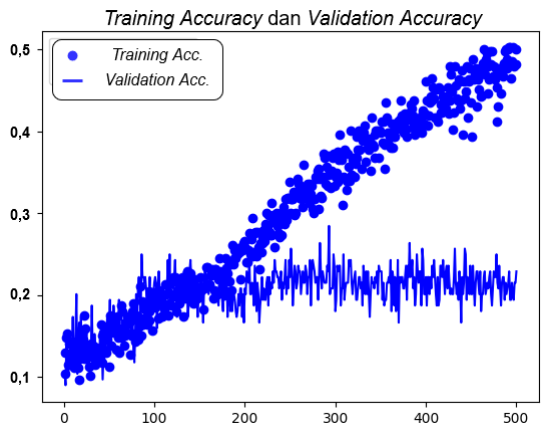
Untuk memastikan model yang dikonstruksi sudah baik, dilakukan percobaan mengubah komposisi data latih-uji menjadi 80:20 dengan satu jenis emosi. Tabel V menunjukkan hasil akurasi pembelajaran per emosi dengan validasi *confusion matrix*. Penjelasan tabel per emosi yaitu netral (Nt), tenang (Tn), senang (Sn), sedih (Sd), marah (Mr), takut (Tk), jijik (Jk), dan kaget (Kgt). Tabel V disajikan berdasarkan setiap himpunan data suara emosi yang digunakan, yaitu RAVDESS, TESS, dan HENLO.



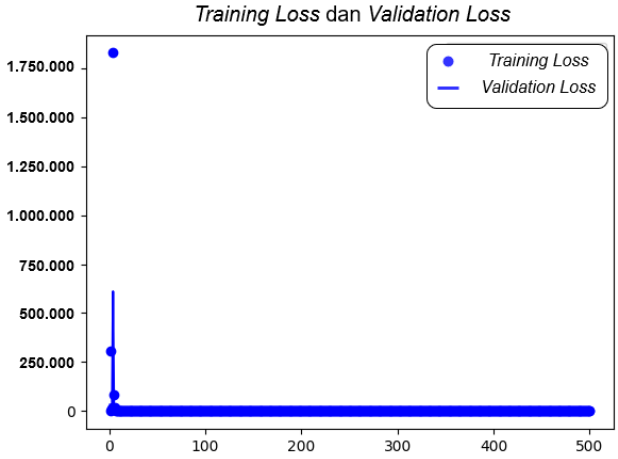
Gbr. 6 Grafik rerata akurasi dengan CNN (komposisi data 80:20).



Gbr. 7 Grafik rerata *loss* dengan CNN (komposisi data 80:20).



Gbr. 8 Grafik rerata akurasi dengan LSTM (komposisi data 80:20).



Gbr. 9 Grafik rerata *loss* dengan LSTM (komposisi data 80:20).

Gbr. 6 dan Gbr. 7 adalah grafik rerata akurasi dan *loss* menggunakan metode pembelajaran CNN. Pada Gbr. 6 terlihat bahwa akurasi mengalami peningkatan seiring dengan jumlah *epoch* yang sudah dilalui dan pada Gbr. 7 terlihat bahwa jumlah *loss* semakin berkurang seiring dengan jumlah *epoch* yang sudah dilalui.

Gbr. 8 dan Gbr. 9 adalah grafik rerata akurasi dan *loss* menggunakan metode pembelajaran LSTM. Pada Gbr. 8 terlihat bahwa akurasi mengalami peningkatan seiring dengan jumlah *epoch* yang sudah dilalui dan pada Gbr. 9 terlihat bahwa jumlah *loss* semakin berkurang seiring dengan jumlah *epoch* yang sudah dilalui.

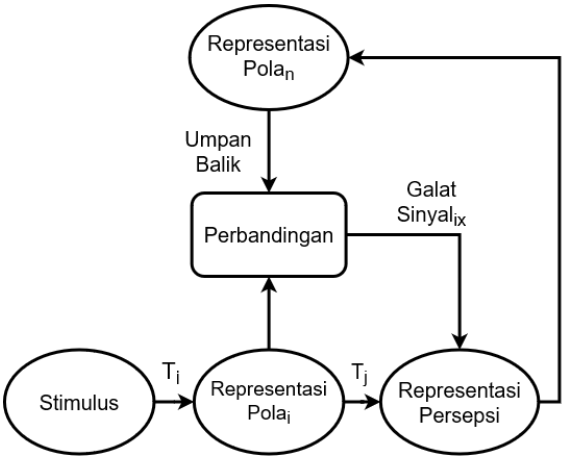
E. Analisis dan Diskusi

Hasil eksperimen menunjukkan perkembangan akurasi pembelajaran dari penelitian lainnya yang menggunakan CNN dan memperoleh rerata akurasi 89,03% [25]. Pada sisi set fitur, penelitian sebelumnya yang menggunakan lima belas fitur sekunder menunjukkan hasil akurasi 78,83% [19]. Dengan kombinasi fitur *cochlegram* dan spektrum Hilbert, diperoleh hasil akurasi pembelajaran sebesar 90,97% untuk CNN dan 80,62% untuk LSTM. Angka akurasi yang tinggi pada penggunaan CNN merupakan keunggulan CNN yang mampu mendeteksi citra dengan cara kerjanya yang mendekati cara kerja reseptor retina manusia.

Unit terkecil dari bunyi yang diproduksi oleh komponen vokal manusia adalah fon, sedangkan fonem adalah unit dari bunyi suara manusia yang membedakan makna-makna di bahasa tertentu. Fonem-fonem dalam suara manusia inilah yang dipindai menjadi spektrum Hilbert dan *cochleagram* untuk mengidentifikasi emosi manusia. Penggunaan fitur *cochleagram* memiliki resolusi lebih baik pada frekuensi rendah daripada skala *mel* yang digunakan dalam *spectrogram*.

Gbr. 10 menunjukkan diagram alir konseptualisasi emosi, yang menurut beberapa ahli neurologi, emosi adalah pengalaman sadar, meskipun merupakan representasi diri dari orang tersebut yang rumit. Representasi tersebut membutuhkan perhatian, metakognisi, dan ingatan yang bekerja, sehingga representasi yang memberikan konten pada pengalaman sadar dapat "diingat" dan dipikirkan [50]–[52]. Emosi memiliki ekspresi penanda atau pengodean yang unik. Masing-masing emosi dapat dipahami, dikategorikan, dan dibedakan dari emosi lain dengan cara yang sama seperti kata-kata yang diartikulasikan oleh pembicara dan dipahami oleh pendengar [53].

Menurut penelitian sebelumnya, sistem *cortical* menghasilkan pengalaman sadar yang sama dari stimulus indrawi yang didapatkan, hanya mendapat *input* berbeda yang menentukan hal yang disadari pada setiap kesempatan [2]. Jika stimulus berupa *input* visual, maka dapat diartikan seseorang sadar



Gbr. 10 Konseptualisasi emosi.

melihat sesuatu. Jika stimulus tersebut adalah *input* pendengaran, maka orang tersebut sadar mendengar sesuatu. Suara ujaran yang diucapkan manusia dapat menimbulkan perasaan yang kuat pada orang-orang tertentu, tetapi mungkin juga tidak jelas pada orang yang lainnya. Dalam *rating* data emosi dalam makalah ini, aktivasi ingatan para psikolog berkorelasi dengan pengalaman emosi atau menggunakan hasil pemikiran kategori emosi tertentu. Proses tersebut melalui beberapa alur: 1) pengodean semantik dari informasi indrawi (pendengaran) menggunakan makna kata yang telah tersimpan di memori; 2) memahami makna kata dari konteks situasi di sekitar suara atau kata yang diucapkan oleh pembicara; dan 3) mengembangkan konteks dalam model-model mental untuk dideskripsikan menurut beberapa model emosi Plutchik.

Jika representasi ini peka terhadap terungkapnya konteks atau pengalaman perseptual langsung, maka hal ini dapat menghasilkan pola dari *input* langsung dan konteks yang berbeda dari pola awal. Umpan balik dari pola berbasis konteks yang dibandingkan dengan pola awal dapat menghasilkan sinyal kesalahan untuk representasi yang mengubah cara konteks diintegrasikan, sehingga menghasilkan pola baru untuk tujuan perbandingan. Atas dasar inilah makalah ini menampilkan identifikasi emosi dari wicara berdasar beberapa model emosi dari Plutchik, yang di-*rating* oleh psikolog, yaitu marah, jijik, takut, bahagia, sedih, netral, dan kaget.

Temuan lain dalam makalah ini yaitu terjadinya jeda waktu beberapa detik bagi psikolog untuk mempertimbangkan pengubahan penilaian deskripsi emosi sampel wicara yang didengarnya. Pengubahan penilaian dan jeda waktu ini terjadi karena psikolog membutuhkan waktu melakukan perbandingan representasi mental dari umpan balik pengalaman maupun konteks informasi tentang label emosi di ingatan dengan konteks para pembicara yang didengar di sampel wicara tersebut. Keraguan mendeskripsikan model emosi terjadi juga karena perasaan adanya perbedaan tingkat dan bentuk emosi saat mendeskripsikan suatu label emosi. Sebagai contoh, emosi takut dalam representasi proses psikologis dinilai memiliki perbedaan secara afektif. Perbedaan ini dapat dinilai sebagai emosi takut karena melihat ke bawah saat naik tangga yang tinggi, tetapi rasa takut akan berbeda jika sedang mengendarai wahana *rollercoaster*. Meskipun sama-sama ketakutan pada ukuran ketinggian, ada emosi lain yang ikut mewarnai deskripsi emosi tersebut. Bentuk emosi takut yang lain yaitu berada di tempat gelap akan memiliki deskripsi berbeda pula jika konteks tempat dan suasana serta pengalaman mental yang pernah dialami berbeda. Hasil eksperimen ini membuka peluang untuk melakukan kajian auditoris selanjutnya terkait dengan konteks konstruksi dinamika sosial dan emosi yang dialami oleh pendengar menggunakan kajian teori selain Plutchik.

V. KESIMPULAN

Rekayasa fitur suara untuk mencari informasi unik emosi dalam suara manusia dilakukan dengan mengimplementasikan CNN dan LSTM, yang masing-masing bekerja dengan baik dalam hal mengenali emosi suara manusia. CNN telah teruji mampu untuk mengenali data emosi melalui fitur suara yang sudah ditransformasi menjadi bentuk dua dimensi.

Pada makalah ini, dilakukan proses identifikasi jenis emosi pada wicara dengan menggabungkan dua fitur spektrum dua dimensi, yaitu fitur *cochleagram* dan spektrum Hilbert, dengan kerangka kerja berbasis konvolusi pada CNN. Hasil eksperimen menunjukkan bahwa kombinasi fitur dua dimensi pada metode CNN memberikan kinerja yang lebih baik dengan total nilai akurasi 90,97% dibandingkan dengan metode LSTM yang memberikan nilai akurasi sebesar 80,62%. Peluang lanjutan yang dapat dilakukan adalah dengan menambahkan jumlah data latih yang lebih banyak dan besar serta penambahan fitur akustik lainnya, khususnya yang berpotensi dalam pengenalan emosi pada wicara.

UCAPAN TERIMA KASIH

Penelitian ini didukung oleh BPPDN Kementrian Riset Teknologi & Pendidikan Tinggi. Terima kasih juga disampaikan kepada Dali Kewara, Randy Anwar Romadhonny, dan Derry Pramono Adi untuk persiapan data dan visualisasi data.

REFERENSI

[1] C. Saint-Georges, M. Chetouani, R. Cassel, F. Apicella, A. Mahdhaoui, F. Muratori, M.-C. Laznik, dan D. Cohen, "Motherese in Interaction: At the Cross-Road of Emotion and Cognition? (A Systematic Review)," *PLoS One*, Vol. 8, No. 10, hal. 1–17, 2013.

[2] R.J. Sternberg dan J.S. Mio, *Cognitive Psychology*, 4th ed. Belmont, USA: Wadsworth, 2005.

[3] J.E. Shackman dan S.D. Pollak, "Experiential Influences on Multimodal Perception of Emotion," *Child Dev.*, Vol. 76, No. 5, hal. 1116–1126, Sep. 2005.

[4] R. Plutchik, "The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact that may Explain Their Complexity and Provide Tools for Clinical Practice," *Am. Sci.*, Vol. 89, No. 4, hal. 344–350, 2001.

[5] M. Gendron, D. Roberson, J.M. van der Vyver, dan L.F. Barrett, "Perceptions of Emotion from Facial Expressions are not Culturally Universal: Evidence from a Remote Culture," *Emotion*, Vol. 14, No. 2, hal. 251–262, 2014.

[6] C.-H. Wu, J.-F. Yeh, dan Z.-J. Chuang, "Emotion Perception and Recognition from Speech," dalam *Affective Information Processing*, J. Tao dan T. Tan, Eds., London, UK: Springer London, 2009, hal. 93–110.

[7] V. Arora, A. Lahiri, dan H. Reetz, "Phonological Feature-based Speech Recognition System for Pronunciation Training in Non-native Language

Learning," *J. Acoust. Soc. Am.*, Vol. 143, No. 1, hal. 98–108, 2018.

[8] B.J. Mohan dan N.R. Babu, "Speech Recognition Using MFCC and DTW," *2014 Int. Conf. Adv. Electr. Eng. (ICAEE 2014)*, 2014, hal. 1-4.

[9] R. Sharma, R.K. Bhukya, dan S.R.M. Prasanna, "Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification," *Speech Commun.*, Vol. 96, hal. 207–224, 2018.

[10] A.R. Avila, S.R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, dan T.H. Falk, "Speech-based Stress Classification Based on Modulation Spectral Features and Convolutional Neural Networks," *Eur. Signal Process. Conf.*, 2019, hal. 1–5.

[11] D.P. Adi, A.B. Gumelar, dan R.P. Arta Meisa, "Interlanguage of Automatic Speech Recognition," *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2019, hal. 88–93.

[12] H.L. Hawkins, T.A. McMullen, A.N. Popper, dan R.R. Fay, *Auditory Computation*, New York, USA: Springer New York, 1996.

[13] K. Venkataramanan dan H.R. Rajamohan, "Emotion Recognition from Speech," *arXiv Prepr. arXiv1912.10458*, hal. 1-14, Des. 2019.

[14] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, dan B. Schmauch, "CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation," *arXiv Prepr. arXiv1802.05630*, hal. 1-5, Feb. 2018.

[15] J. Zhao, X. Mao, dan L. Chen, "Speech Emotion Recognition using Deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, Vol. 47, hal. 312–323, Jan. 2019.

[16] S. Basu, J. Chakraborty, dan M. Aftabuddin, "Emotion Recognition from Speech Using Convolutional Neural Network with Recurrent Neural Network Architecture," *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, hal. 333–336.

[17] M. Slaney dan R.F. Lyon, "On the Importance of Time - A Temporal Representation of Sound," dalam *Visual Representations of Speech Signals*, M. Cooke, S. Beet, dan M. Crawford, Eds., Hoboken, USA: John Wiley & Sons Ltd, 1993, hal. 95–116.

[18] V.K. Rai dan A.R. Mohanty, "Bearing Fault Diagnosis using FFT of Intrinsic Mode Functions in Hilbert–Huang Transform," *Mech. Syst. Signal Process.*, Vol. 21, No. 6, hal. 2607–2615, Agt. 2007.

[19] A.B. Gumelar, A. Kurniawan, A.G. Sooai, M.H. Purnomo, E.M. Yuniarno, I. Sugiarto, A. Widodo, A.A. Kristanto, dan T.M. Fahrudin, "Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks," *2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)*, 2019, hal. 1–8.

[20] Y.K. Muthusamy, R.A. Cole, dan M. Slaney, "Speaker-independent Vowel Recognition: Spectrograms Versus Cochleagrams," *International Conference on Acoustics, Speech, and Signal Processing*, 1990, hal. 533–536.

[21] S. Sandoval, P.L. de Leon, dan J.M. Liss, "Hilbert Spectral Analysis of Vowels using Intrinsic Mode Functions," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, hal. 569–575.

[22] S. Camacho, D. Renza, dan D.M. Ballesteros L., "A Semi-supervised Speaker Identification Method for Audio Forensics Using Cochleagrams," dalam *Applied Computer Sciences in Engineering: 4th Workshop on Engineering Applications (WEA 2017)*, J.C. Figueroa-García, E.R. López-Santana, J.L. Villa-Ramírez, dan R. Ferro-Escobar, Eds., Cham, Switzerland: Springer, 2017, hal. 55–64.

[23] B. Gao, W.L. Woo, dan L.C. Khor, "Cochleagram-based Audio Pattern Separation using Two-Dimensional Non-negative Matrix Factorization with Automatic Sparsity Adaptation," *J. Acoust. Soc. Am.*, Vol. 135, No. 3, hal. 1171–1185, Mar. 2014.

[24] X.L. Zhang dan D.L. Wang, "Boosted Deep Neural Networks and Multi-Resolution Cochleagram Features for Voice Activity Detection," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2014, hal. 1534–1538.

[25] R.V. Sharan dan T.J. Moir, "Cochleagram Image Feature for Improved Robustness in Sound Recognition," *Int. Conf. Digit. Signal Process. DSP*, 2015, hal. 441–444.

[26] C. Darwin, "Computational Auditory Scene Analysis: Principles, Algorithms and Applications," *J. Acoust. Soc. Am.*, Vol. 124, No. 1, hal. 13–13, Jul. 2008.

[27] M.K.I. Molla dan K. Hirose, "Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum," *IEEE Trans. Audio, Speech Lang. Process.*, Vol. 15, No. 3, hal. 893–900, Mar. 2007.

[28] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, dan M. Allerhand, "Complex Sounds and Auditory Images," *Proceedings of the 9th International Symposium on Hearing*, 1991, hal. 429–446.

[29] H. Yin, V. Hohmann, dan C. Nadeu, "Acoustic Features for Speech Recognition Based on Gammatone Filterbank and Instantaneous Frequency," *Speech Commun.*, Vol. 53, No. 5, hal. 707–715, Mei 2011.

[30] M. Russo, M. Stella, M. Sikora, dan V. Pekić, "Robust Cochlear-Model-Based Speech Recognition," *Computers*, Vol. 8, No. 1, hal. 1-15, Jan. 2019.

[31] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Apple Compupter, Inc., Cupertino, USA, Technical Report, 1993.

[32] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.-C. Yen, C.C. Tung, dan H.H. Liu, "The Empirical Mode Decomposition and the Hubert Spectrum for Non-linear and Non-stationary Time Series Analysis," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, Vol. 454, No. 1971, hal. 903–995, 1998.

[33] H. Huang dan X.-X. Chen, "Speech Formant Frequency Estimation based on Hilbert-Huang Transform," *Zhejiang Daxue Xuebao (Gongxue Ban)/Journal Zhejiang Univ. (Engineering Sci.*, Vol. 40, hal. 1926–1930, 2006.

[34] H. Huang dan J. Pan, "Speech Pitch Determination based on Hilbert-Huang Transform," *Signal Processing*, Vol. 86, No. 4, hal. 792–803, Apr. 2006.

[35] X. Li dan X. Li, "Speech Emotion Recognition Using Novel HHT-TEO Based Features," *J. Comput.*, Vol. 6, No. 5, hal. 989-998, Mei 2011.

[36] A.B. Gumelar, M.H. Purnomo, E.M. Yuniarno, dan I. Sugiarto, "Spectral Analysis of Familiar Human Voice Based On Hilbert-Huang Transform," *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, 2018, hal. 311–316.

[37] N.E. Huang dan S.S.P. Shen, *Hilbert-Huang Transform and Its Applications*, Singapore, Singapore: World Scientific, 2005.

[38] S.K. Phan dan C. Chen, "Big Data and Monitoring the Grid," dalam *The Power Grid*, B.W. D'Andrade, Eds., Cambridge, USA: Academic Press, 2017, hal. 253–285.

[39] S.R. Livingstone dan F.A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," *PLoS One*, Vol. 13, No. 5, hal. 1-35, 2018.

[40] K. Dupuis dan M.K. Pichora-Fuller, "Toronto Emotional Speech Set (TESS)," Scholars Portal Dataverse, 2010.

[41] X.-L. Zhang dan Ji Wu, "Deep Belief Networks Based Voice Activity Detection," *IEEE Trans. Audio. Speech. Lang. Processing*, Vol. 21, No. 4, hal. 697–710, Apr. 2013.

[42] L. Mateju, P. Cerva, dan J. Zdansky, "Study on the Use of Deep Neural Networks for Speech Activity Detection in Broadcast Recordings," *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications*, 2016, hal. 45–51.

[43] S. Thomas, S. Ganapathy, G. Saon, dan H. Soltau, "Analyzing Convolutional Neural Networks for Speech Activity Detection in Mismatched Acoustic Conditions," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, hal. 2519–2523.

[44] L. Deng dan J. Platt, "Ensemble Deep Learning for Speech Recognition," *Proc. Interspeech*, 2014, hal. 1-5.

[45] T. Xu, H. Li, H. Zhang, dan X. Zhang, "Improve Data Utilization with Two-stage Learning in CNN-LSTM-based Voice Activity Detection," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, hal. 1185–1189.

[46] T.N. Sainath, O. Vinyals, A. Senior, dan H. Sak, "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, hal. 4580–4584.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, dan D. Cournapeau, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, Vol. 12, hal. 2825–2830, 2011.

[48] S. Hochreiter dan J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, Vol. 9, No. 8, hal. 1735–1780, Nov. 1997.

[49] F.A. Gers, "Learning to Forget: Continual Prediction with LSTM," *9th International Conference on Artificial Neural Networks (ICANN '99)*, 1999, hal. 850–855.

[50] J. LeDoux dan A.R. Damasio, "Emotions and Feelings," dalam *Principles of Neural Science*, 5th ed., E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum, dan A.J. Hudspeth, Eds. New York, USA: McGraw-Hill, 2013, hal. 1079–1094.

[51] J.E. LeDoux dan R. Brown, "A Higher-order Theory of Emotional Consciousness," *Proc. Natl. Acad. Sci. of USA*, Vol. 114, No. 10, hal. E2016–E2025, Mar. 2017.

[52] J.E. LeDoux, "Semantics, Surplus Meaning, and the Science of Fear," *Trends Cogn. Sci.*, Vol. 21, No. 5, hal. 303–306, Mei 2017.

[53] K.R. Scherer, E. Clark-Polner, dan M. Mortillaro, "In the Eye of the Beholder? Universality and Cultural Specificity in the Expression and Perception of Emotion," *Int. J. Psychol.*, Vol. 46, No. 6, hal. 401–435, Des. 2011.