Ulasan: Pengenalan Emosi Melalui Suara

Rio Galang Jati Respati
Program Studi Informatika – Program Sarjana
Universitas Islam Indonesia
Yogyakarta, Indonesia
17523118@students.uii.ac.id

Arrie Kurniawardhani
Jurusan Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
arrie.kurniawardhani@uii.ac.id

Abstract*—Penggunaan suara dalam data science telah banyak dilakukan oleh beberapa peneliti seperti mendeteksi suara hewan, ataupun objek tertentu. Namun, penelitian tentang emosi manusia sedang gencar sekali, dan salah satunya adalah mendeteksi emosi manusia melalui suara. Dewasa ini pendeteksian emosi tidak hanya digunakan untuk riset di bidang akademik seperti psikologi, neuroscience, psikiater, ilmu kognitif dan lainya. Tetapi juga diaplikasikan secara praktis seperti call centre, gaming industry, bidang medis dan lainya. Penelitian ini menjelaskan tentang perkembangan pendeteksian emosi melalui suara dari tahun 2017 hingga 2019. Perkembangan dalam mendeteksi sebuah sinyal suara berdasarkan emosi berkembang cukup pesat. Salah satu penelitian dalam mendeteksi sebuah sinyal suara dapat mencapai akurasi yang cukup tinggi, dengan persentase 95.33%.. Untuk penggunakan model (SVM) dan (CNN) sudah cukup bisa dikatakan layak untuk digunakan sebagai model yang tepat untuk medeteksi emosi melalui suara.

***Kata kunci*— emosi, suara, deteksi, kajian pustaka**

I. PENDAHULUAN

Penggunaan suara dalam *data science* telah banyak dilakukan oleh beberapa peneliti seperti mendeteksi suara hewan, ataupun objek tertentu. Namun, penelitian tentang emosi manusia sedang gencar sekali, dan salah satunya adalah mendeteksi emosi manusia melalui suara [1]. Dewasa ini pendeteksian emosi tidak hanya digunakan untuk riset di bidang akademik seperti psikologi, *neuroscience*, psikiater, ilmu kognitif dan lainya. Tetapi juga diaplikasikan secara praktis seperti *call centre, gaming industry*, bidang medis dan lainya [1]. Emosi seseorang memang belum bisa ditentukan secara langsung oleh suara, karena deteksi emosi biasanya dilakukan oleh keduanya yaitu *gimmick* dari bentuk wajah dan sinyal dari suara seseorang, namun setidaknya emosi seseorang dapat diketahui dengan sinyal suara, walaupun akurasi deteksi yang dikeluarkan rendah. Karakteristik dan kepribadian manusia memiliki perbedaan, baik itu wajah, tubuh, rambut, dan suara. Contoh perbedaan signifikan yang dimiliki manusia yaitu suara karena perbedaan suara tersebut dapat membuat komputer sulit mengetahui emosi dari suara yang dikeluarkan orang tersebut [2]. Apabila kemungkinan tersebut terjadi, Jianfeng Zhao [2] beranggapan bahwa suara yang dilontarkan dengan nada tinggi merupakan suara dari seseorang yang sedang berbicara dengan normal, sebaliknya ada juga yang beranggapan bahwa suara yang dilontarkan dengan nada tinggi merupakan suara dari seseorang yang sedang marah. Studi literatur ini bertujuan untuk mengetahui perkembangan yang terjadi dalam pendeteksian emosi melalui suara. Perkembangan yang diamati mulai dari *database*, metode dalam ekstraksi fitur, hingga model pengklasifikasian yang digunakan.

II. SUMBER METODE

Literatur diperoleh melalui beberapa portal edukasi yang sudah terpercaya dan mudah dilacak kembali seperti Google Scholar. Fokus pencarian literatur menggunakan *keyword emotion, speech, dan recognition*. Batasan dalam pengumpulan literatur ini berdasarkan umur literatur tersebut maksimal 4 tahun dari tanggal publikasi, lingkup isinya terkait Analisis, Metode, dan jenis literatur yang digunakan yaitu jurnal.

III. HASIL DAN DISKUSI

A. Database Suara

Dario Bartero [4] menyatakan bahwa dataset yang digunakan berasal dari TEDLIUM v2 corpus. *Database* terdapat lebih dari 5000 data yang memuat beberapa emosi yaitu sedih, marah, senang, netral, dan garbage. Data dari *database* yang digunakan, sekitar 90% berasal dari murid di dalam research group, sisanya berasal dari data crowdsourcing dari Amazon Mechanical Turk. Dari 5000 lebih data tersebut, terdiri dari 877 data untuk kelas “sedih”, 771 data untuk kelas “marah”, 3498 data untuk kelas “senang”, dan sisanya adalah data untuk kelas “netral” atau “garbage” [4]. Penelitian lain yang dilakukan oleh Nancy Semwal [5], menyebutkan bahwa dataset yang digunakan berasal dari dua tempat yang berbeda, yaitu Berlin *Database* of Emotional Speech (EmoDB) dan RML Emotion *Database* (RED). EmoDB berisikan 535 data dari tujuh emosi yaitu marah, gelisah, bosan, jijik, senang, netral, dan sedih. Data tersebut menggunakan bahasa jerman dan disuarakan oleh lima laki laki dan lima perempuan. Setiap penyuara berbicara sepuluh kalimat yang sama dari enam

emosi yang berbeda. Namun, di dalam penelitian tersebut emosi jijik dihapus sehingga tersisa data sejumlah 489. Pada data RED berisikan 720 data yang terdiri dari 6 emosi yang berbeda dan setiap emosi terbagi rata sejumlah 120 data. Emosi yang terdapat di dalam *database* tersebut antara lain marah, jijik, takut, senang, sedih, dan terkejut. *Database* RED terdiri dari berbagai bahasa, yaitu inggris, mandarin, urdu. persian, italian dan punjabi.

Penelitian yang dilakukan oleh Manas Jain [1] menggunakan dua *database* yang berbeda, yaitu Linguistic Data Consortium (LDC), dan UGA. Kedua *database* tersebut memiliki jumlah data yang sama, yaitu 100 data. Emosi yang terdapat pada *database* terdiri dari sedih, senang, marah, dan netral [1]. Udit Jain [3] melakukan penelitian dengan mengambil dataset yang berasal dari platform yang cukup besar yaitu Youtube. Dataset tersebut diunduh dari rekaman suara 20 artis yang berbeda terdiri dari 10 laki laki dan 10 perempuan. Rekaman suara tersebut berisi suara artis yang mengucapkan beberapa emosi terdiri dari senang, sedih, marah, dan netral. Setiap emosi pada rekaman tersebut berdurasi dua detik dan berformat .wav. Jumlah data yang didapatkan sebesar 400 audio file yang terdiri dari 200 suara laki laki dan 200 suara perempuan. Pembagian emosi per *database* cukup rata, tiap emosi terdapat 50 file [3].

Jianfeng Zhao [2], menggunakan dua *database* yang berbeda, yaitu EmoDB dan Interactive Emotional Speech Dataset (IEMOCAP). *Database* EmoDB memiliki 535 files dan tujuh emosi yang berbeda. IEMOCAP memiliki 1150 data yang disuarakan oleh lima penyuara (laki laki dan perempuan). Database tersebut memiliki tujuh emosi yang terdiri dari marah, excited, frustasi, senang, netral, sedih, dan terkejut. Database secara detail terbagi atas 71 suara termasuk dalam kelas "marah", 178 suara termasuk dalam kelas "excited", 271 suara termasuk dalam kelas "frustasi", 31 suara termasuk dalam kelas "senang", 328 suara termasuk dalam kelas "netral", 265 suara termasuk dalam kelas "sedih", dan 6 suara termasuk dalam kelas "terkejut" [2].

Suraj Tripathi [6], telah melakukan penelitian dengan mengambil dataset yang berasal dari University of Southern California's-Interactive Emotional Motion Capture (USC-IEMOCAP). Isi dari database tersebut adalah percakapan antara laki-laki dan perempuan. Topik dari percakapan tersebut menghasilkan beberapa emosi, yaitu marah, senang, netral, dan sedih. Satu emosi didapatkan dari potongan percakapan yang panjang antara 3-15 detik. Database secara detail terbagi atas 48.8% suara termasuk dalam kelas "netral", 12.3% suara termasuk dalam kelas "senang", 26.9% suara termasuk dalam kelas "sedih", dan 12% suara termasuk dalam kelas "marah" [6].

B. Ekstraksi Fitur

Ekstraksi fitur yang digunakan oleh Dario Bertero [1] & Zhao Jianfeng [3] menggunakan ekstraksi fitur pada layer konvolusi dan tidak menggunakan metode tambahan. Nancy Semwal dalam penelitiannya menyatakan bahwa metode Low Level Descriptors (LLDs) digunakan karena tepat dalam mengekstraksi dari sinyal suara pada percakapan yang pendek. Penelitian lain yang dilakukan oleh Manas Jain, dalam mengekstraksi fitur menggunakan dua metode, yaitu *Mel-Frequency cepstral coefficients* (MFCCs) dan *Linear Prediction Cepstral Coefficient* (LPCC). MFCCs merupakan metode yang umum digunakan dalam mengekstraksi fitur. Perhitungan metode tersebut didasarkan pada karakter telinga manusia sehingga perhitungannya mirip sistem pendengaran manusia. LPCC merupakan metode yang digunakan untuk mengulas setiap *envelope* yang berisi sinyal suara percakapan yang telah terkompresi. LPCC menggunakan informasi dari model linear productive. Model tersebut menganalisis sinyal yang masuk dengan mengestimasi forman atau pita frekuensi yang telah ditingkatkan. Model tersebut juga menghapus efek yang muncul dari sinyal dan mengestimasi intensitas frekuensi dari sinyal yang tersisa. Ekstraksi fitur yang dilakukan oleh Suraj Tripati [6] tidak hanya bergantung pada layer konvolusi, bahkan mencoba metode MFCCs dengan menggunakan phyton package yang bernama librosa. Hal ini juga dilakukan oleh Udit Jain [3] yang menggunakan tiga metode ekstraksi fitur antara lain metode MFCCs dengan akurasi yang tinggi mampu mengklasifikasi frekuensi senstivitas persepsi manusia, kemudian metode LPCC merupakan metode perhitungan koefisien dimana setiap sampel suara pada sumbu waktu x saat ini dapat diproyeksikan sebagai kombinasi linier dari sampel suara sebelumnya yang dimodelkan menggunakan filter *digital all-pole*, dan metode *Formant Frequency* (FF) merupakan metode dengan konsentrasi energi akustik dalam rentang frekuensi tertentu dapat didefinisikan sebagai forman gelombang wicara tertentu dan mewakili kekuatan sinyal wicara yang diperoleh pada frekuensi tertentu.

C. Klasifikasi

Menurut Dario Bertero [4] menggunakan *Convolutional Neural Network* (CNN) sebagai model klasifikasi, Dario Bertero hanya menggunakan satu lapisan saat konvolusi, lalu model klasifikasi tersebut membuat waktu evaluasi menjadi lebih cepat dibandingkan menggunakan struktur dua lapis. Lapisan konvolusi tersebut mengekstraksi setiap fitur pada frame dan mengevaluasi perbedaan dari frame yang overlapping. Kemudian setelah klasifikasi pada lapisan konvolusi terdapat proses *max-pooling*. *Max-pooling* mengizinkan kita untuk dapat memilih kontribusi dari setiap *frame* yang paling signifikan, dan untuk mengkombinasi keseluruhan menjadi vektor yang sudah ditetapkan ukurannya. Pada lapisan *fully connected* (FC) yang memiliki lapisan sebanyak dua ratus dan diikuti oleh lapisan softmax yang melakukan klasifikasi yang sebenarnya. Proses metode berbasis deeplearning disebut sebagi "magical boxes". Selain melakukan klasifikasi proses deeplearning ini digunakan untuk merepresentasikan fitur suara dan untuk memverifikasi apa yang terjadi di dalam "magical boxes" tersebut. Dario Bertero membandingkan akurasi yang dimiliki jika menggunakan *Support Vector Machine* (SVM) dan *Convolutional Neural Network* (CNN) [4].

Menurut Nancy Semwal [5] dalam menggunakan *Support Vector Machine* (SVM) sebagai model klasifikasi. *Support Vector Machine* (SVM) dapat mentrasfer kumpulan fitur-fitur yang ada di dalam vektor, lalu setiap vektor akan diberikan

sebuah tanda untuk masuk ke dalam satu atau dua kategori. Proses tersebut dilakukan dengan menggunakan kernel functions. (kernel functions adalah). Kemudian *Support Vector Machine* (SVM) sangat efektif ketika digunakan dalam sebuah kasus fitur dimensi yang lebih tinggi daripada jumlah kasus pada sampel. Lalu pada penelitian kali ini Nancy Semwal menggunakan implementasi dari *library* bernama LIBSVM [5].

Menurut Manas Jain [1] menggunakan model *Support Vector Machine* (SVM) sebagai model klasifikasi. Lalu Manas Jain berpendapat bahwa model *Support Vector Machine* (SVM) ini sangat simpel dan efisien untuk mengklasifikasi dan pengenalan pola menggunakan algoritma klasifikasi. Ada dua jenis *Support Vector Machine* (SVM) yaitu linear dan nonlinear. Manas Jain menggunakan LIBSVM sebagai *library* yang digunakan untuk klasifikasi ketika menggunakan *Support Vector Machine* (SVM). Manas Jain mengatakan ketika model sudah selesai disiapkan implementasi klasifikasi dan pengenalan pola akan menjadi sangat mudah untuk mempredeksi emosi dari dataset uji. [1]

Menurut Udit Jain [3] dalam menggunakan SVM sebagai model klasifikasi. Udit Jain [3] melakukan segmentasi data tersebut menggunakan k-fold Cross-Validation untuk masuk ke training *database* dan testing *database*. Setelah melalui segmentasi data menggunakan k-fold Cross-Validation, data tersebut akan dimasukkan ke dalam model *Support Vector Machine* (SVM). Berbeda dari *Support Vector Machine* yang biasanya, Udit Jain [3] menggunakan Cubic *Support Vector Machine* (SVM) sebagai model klasifikasi, karena Cubic *Support Vector Machine* (SVM) memiliki waktu komputasi lebih sedikit 10 detik dari model *Support Vector Machine* pada umumnya [3].

Menurut Suraj Tripathi [6] dalam menggunakan model *Convolutional Neural Network* (CNN) pada layer konvolusinya terdapat empat paralel. Pada setiap konvolusi menggunakan 200 kernel di setiap paralel. Untuk mencegah kurang optimalisasi ketika memilih kernel Suraj Tripathi memutuskan untuk membuat ukuran kernel yang ukurannya berbeda-beda. Selanjutnya setelah dilakukan konvolusi pada layer pertama hasil konvolusi tersebut akan dikirim kepada lapisan max-pooling. Kemudian setelah diterima oleh lapisan max-pooling hasil konvolusi tersebut dikirim kepada lapisan *Fully connected* (FC), setelah di terima oleh lapisan *Fully connected* (FC) hasil konvolusi kembali dikirim kepada lapisan terakhir yaitu lapisan softmax, seperti itu proses pembagian untuk mendeteksi empat emosi yang berbeda [6].

Menurut Jianfeng Zhao [2] dalam menggunakan model *Convolutional Neural Network* (CNN) tedapat dua macam yaitu *1Dimensional Convolutional Neural Network Long Short – Term Memory* (1D CNN LSTM) dan *2Dimensional Convolutional Neural Network Long Short – Term Memory* (2D CNN LSTM). *1Dimensional Convolutional Neural Network Long Short – Term Memory* disusun dengan Local Feature Larning Blocks (LFLBs) sebanyak empat lapis, lalu lapisan *Long Short-Term Memory* sebanyak satu lapis, dan lapisan *Fully connected* (FC) sebanyak satu lapis. *1Dimensional Convolutional Neural Network Long Short – Term Memory* (1D CNN LSTM) dibuat untuk mempelajari fitur dari audio yang bersifat raw, maka dari itu, konvolusi dan pooling kernel di tiap *Local Feature Learning Block* (LFLB) hanya memiliki satu dimensi. Oleh karena itu, ketika ada suara yang direpresentasikan sebagai satu dimensional vektor, maka suara tersebut akan dimasukkan ke model 1 Dimensional *Convolutional Neural Network Long Short – Term Memory* (1D CNN LSTM), lalu fitur suara yang telah masuk tersebut akan dipelajari oleh *Local Feature Learning Blocks* (LFLBs). Setelah suara tersebut dipelajari, fitur yang terdapat dari suara tersebut akan dimasukkan ke lapisan *Long Short-Term Memory* (LSTM), lalu fitur tersebut akan dipindahkan ke lapisan *Fully connected* (FC). *2Dimensional Convolutional Neural Network Long Short – Term Memory* (2D CNN LSTM) mempunyai struktur yang sama dengan 1 Dimensional *Convolutional Neural Network Long Short – Term Memory* (1D CNN LSTM). Perbedaannya terletak pada bentuk dari konvolusi dan pooling kernel pada *Local Feature Learning Blocks* (LFLBs) adalah dua dimensi. *2Dimensional Convolutional Neural Network Long Short – Term Memory* (2D CNN LSTM) dibuat untuk mempelajari fitur high-level oleh emosi yang berasal dari spektogram log-mel. Oleh karena itu, ketika terdapat spektogram log-mel dalam bentuk matriks dimasukkan ke dalam *2Dimensional Convolutional Neural Network Long Short – Term Memory* (2D CNN LSTM), fitur lokasl akan dipelajari oleh empat *Local Feature Learning Blocks* (LFLBs). Setelah fitur tersebut dipelajari oleh *Local Feature Learning Blocks* (LFLBs) fitur tersebut akan dibentuk ulang menjadi sekuensi yang sementara dan dipindahkan ke lapisan *Long Short-Term Memory* (LSTM). Lapisan *Fully connected* (FC) akan menggeneralisasi fitur-fitur yang telah dikeluarkan oleh *Long Short-Term Memory* (LSTM), lalu
lapisan Softmax akan membuat prediksi berdasarkan fitur tersebut [2].

D. Akurasi

Dorio Bertero [4] mendapatkan akurasi di angka 66.1% pada model CNN dan 63.0% pada model SVM. Untuk lebih detail di tiap emosinya, akurasi saat menggunakan CNN adalah sebagai berikut, pada emosi “marah” mendapatkan akurasi 70%, emosi senang mendapatkan akurasi 58.6%, sedangkan emosi sedih mendapatkan akurasi 69.1%, lalu dibandingkan saat menggunakan SVM, hasilnya sebagai berikut, pada emosi marah mendapatkan akurasi 60.4%, pada emosi senang mendapatkan akurasi 52.2%, dan pada emosi sedih menghasilkan akurasi yang cukup baik di 76.4%. Akurasi saat senang terbilang rendah karena data saat emosi senang sangat banyak jika dibandingkan yang lain. Hal ini kemungkinan terjadi karena emosi saat “netral” tidak berbeda jauh dengan “senang”, dan membuat yang sebenarnya “netral” masuk ke kelas “senang” [4].

No	Model	Emosi	Akurasi (%)
1	CNN	Marah	70
2	CNN	Senang	58.6
3	CNN	Sedih	69.1

4	SVM	Marah	60.4
5	SVM	Senang	52.2
6	SVM	Sedih	76.4
Rata rata (CNN)			66.1
Rata rata (SVM)			63

TABEL 1. AKURASI DARI MODEL CNN DAN SVM YANG DIBUAT OLEH DORIO BERTERO

Nancy Semwal [5] di tahun yang sama mendapatkan akurasi yang cukup tinggi di kedua *database* yang ia gunakan. Beliau mendapatkan akurasi rata rata 80% ketika menggunakan *database* EmoDB, lalu 73% ketika menggunakan *database* RED. Akurasi lebih detail ketika menggunakan *database* EmoDB adalah sebagai berikut, SVM dapat menebak 107 file yang benar dari 127 file ketika emosi "marah", 59 file dari 69 file ketika emosi "gelisah", 67 file dari 81 file ketika emosi "bosan", 48 file dari 71 file ketika senang, 59 file dari 79 file saat "netral", dan 58 file dari 62 file saat "sedih". Sedangkan akurasi lebih detail ketika menggunakan *database* RED adalah sebagai berikut, 98 file dari 120 file untuk emosi "marah", 84 file dari 120 file untuk emosi "jijik", 77 file dari 120 file untuk emosi "takut", 75 file dari 120 file untuk emosi "senang", 97 file dari 120 file untuk emosi "sedih", 97 file dari 120 file untuk emosi "terkejut" [5].

No	*Database*	Emosi	Akurasi (%)
1	EmoDB	Marah	84.2
2	EmoDB	Gelisah	85.5
3	EmoDB	Bosan	82.7
4	EmoDB	Senang	67.6
5	EmoDB	Netral	74.6
6	EmoDB	Sedih	93.5
7	RED	Marah	81.6
8	RED	Jijik	70
9	RED	Takut	64.1
10	RED	Senang	62.5
11	RED	Sedih	80.3
12	RED	Terkejut	80.3
Rata rata (*Database* EmoDB)			80
Rata rata (*Database* RED)			73

TABEL 2. AKURASI DARI MODEL SVM YANG DIBUAT OLEH NANCY SEMWAL

Manas Jain [1] mendapatkan akurasi rata rata 79.22% ketika menggunakan LDC sebagai database, lalu mendapatkan akurasi rata rata 50.38% ketika menggunakan UGA sebagai database, akurasi tersebut menggunakan MFCCs sebagai metode ekstraksi fitur. Detail akurasi untuk penggunaan LDC sebagai database adalah sebagai berikut, 98.52% dalam emosi "senang", 63.63% dalam emosi "sedih", 71.42% dalam emosi "marah", 83.33% dalam emosi "takut". Kemudian berlanjut jika menggunakan *database* dari UGA, detail akurasi sebagai berikut, 42.85% dalam emosi "senang", 54.54% dalam emosi "sedih", 66.66% dalam emosi "marah", dan 37.50% dalam emosi "takut" [1].

No	*Database*	Emosi	Akurasi (%)
1	LDC	Senang	98.52
2	LDC	Sedih	63.63
3	LDC	Marah	71.42
4	LDC	Takut	83.33
5	UGA	Senang	42.85
6	UGA	Sedih	54.54
7	UGA	Marah	66.66
8	UGA	Takut	37.50
Rata rata (LDC)			79.22
Rata rata (UGA)			50.38

TABEL 3. AKURASI DARI MODEL SVM KETIKA MENGGUNAKAN MFCC YANG DIBUAT OLEH MANAS JAIN

Manas Jain [1] juga membandingkan akurasi ketika menggunakan dua metode ekstraksi yang berbeda yaitu MFCCs dan LPCC. Manas Jain [1] menggunakan LDC database sebagai database ketika membandingkan kedua metode ekstraksi fitur tersebut. Manas Jain mendapatkan detail akurasi sebagai berikut, 70.94% dalam emosi "senang", 71.32% dalam emosi "sedih", 85.65% dalam emosi "marah", dan 64.59% dalam emosi "takut" [1].

No	*Database*	Emosi	Akurasi (%)
1	LDC	Senang	70.94
2	LDC	Sedih	71.32
3	LDC	Marah	85.65
4	LDC	Takut	64.59
Rata rata (LDC)			73.12

TABEL 4. AKURASI DARI MODEL SVM KETIKA MENGGUNAKAN LPCC YANG DIBUAT OLEH MANAS JAIN

Udit Jain [3] mendapatkan akurasi yang bagus saat melakukan penelitian pada tahun 2018. Beliau membandingkan akurasi saat menggunakan metode ekstraksi fitur yang berbeda, ketika menggunakan MFCCs mendapatkan akurasi 22%, lalu ketika menggunakan LPCC mendapatkan akurasi di 25%. Kemudian beliau menggabungkan kedua metode tersebut dan menghasilkan akurasi yang cukup jauh di angka 56%. Lalu kemudian beliau menggabungkan lagi dengan metode FF dan hasilnya membuat akurasi menjadi lebih tinggi di angka 81% [3].

No	Metode Ekstraksi	Rata rata akurasi (%)
1	MFCCs	22
2	LPCC	25
3	MFCCs + LPCC	56
4	MFCCs + LPCC + FF	81

TABEL 5. AKURASI DARI METODE EKSTRAKSI FITUR OLEH UDIT JAIN

Jianfeng Zhao [2] mendapatkan akurasi diatas 80% dari kedua *database* yang digunakan. Beliau juga menggunakan dua klasifikasi yang berbeda yaitu CNN 1D dan CNN 2D. Ketika menggunakan EmoDB dengan model CNN 1D, beliau mendapatkan rata rata akurasi di 92%. Untuk detail akurasi sebagai berikut, 92.91% ketika emosi "marah", 98.77% ketika emosi "bosan", 76.09% ketika emosi "jijik", 94.2% ketika emosi "takut", 69.01% ketika emosi "senang", 78.48 % ketika emosi "netral", 88.71% ketika emosi "sedih". Hasil yang didapatkan sedikit lebih tinggi jika dibandingkan menggunakan model CNN 2D, beliau mendapatkan rata rata akurasi di 95.33%. Untuk detail akurasi sebagai berikut, 100% saat emosi "marah", 97.53% saat emosi "bosan", 86.96% saat emosi "jijik", 97.1% saat emosi "takut", 91.55% saat emosi "senang", 93.67% saat emosi "netral", 98.39% saat emosi "sedih". Jika menggunakan *database* IEMOCAP dengan menggunakan model CNN 1D, mendapatkan rata rata akurasi sebesar 79.72%. Untuk detail akurasinya sebagai berikut, 90.14% saat emosi "marah", 83.71% saat emosi "excited", 78.6% saat emosi "frustasi", 41.94% saat emosi "senang", 68.29% saat emosi "netral", dan 95.85% saat emosi "sedih". Sama seperti saat menggunakan EmoDB, di *database* ini juga memiliki akurasi yang sedikit lebih tinggi saat menggunakan model CNN 2D. Akurasi yang didapat naik hingga 85.58%. Detail dari akurasi sebagai berikut, 84.51% saat emosi "marah", 88.2% saat emosi "excited", 75.65% saat emosi "frustasi", 41.94% saat emosi "senang", 89.94% saat emosi "netral", 96.98% saat emosi "sedih" [2].

No	Model CNN	*Database*	Emosi	Akurasi (%)
1	1D	EmoDB	Marah	92.91
2	1D	EmoDB	Bosan	98.77
3	1D	EmoDB	Jijik	76.09
4	1D	EmoDB	Takut	94.2
5	1D	EmoDB	Senang	69.01
6	1D	EmoDB	Netral	78.84
7	1D	EmoDB	Sedih	88.71
8	1D	IEMOCAP	Marah	90.14
9	1D	IEMOCAP	Excited	83.71
10	1D	IEMOCAP	Frustasi	78.6
11	1D	IEMOCAP	Seneng	41.94
12	1D	IEMOCAP	Netral	68.29
13	1D	IEMOCAP	Sedih	95.85
14	2D	EmoDB	Marah	100
15	2D	EmoDB	Bosan	97.53
16	2D	EmoDB	Jijik	86.96
17	2D	EmoDB	Takut	97.1
18	2D	EmoDB	Senang	91.55
19	2D	EmoDB	Netral	93.67
20	2D	EmoDB	Sedih	98.39
21	2D	IEMOCAP	Marah	84.51
22	2D	IEMOCAP	Excited	88.2
23	2D	IEMOCAP	Frustasi	75.65
24	2D	IEMOCAP	Seneng	41.94
25	2D	IEMOCAP	Netral	89.94
26	2D	IEMOCAP	Sedih	96.98
Rata-rata (1D EmoDB)				92
Rata-rata (1D IEMOCAP)				79.72
Rata-rata (2D EmoDB)				95.33
Rata-rata (2D IEMOCAP)				85.58

TABEL 6. AKURASI DARI MODEL CNN YANG DIBUAT OLEH JIANFENG ZHAO

Suraj Tripathi [6] mendapatkan rata rata akurasi sebesar 76.1%. Untuk detail akurasi adalah sebagai berikut, 81.30% saat emosi "netral", 49.24% saat emosi "senang", 84.06% saat emosi sedih, 63.41% saat emosi "marah" [6].

No	Emosi	Akurasi (%)
1	Netral	81.3
2	Senang	49.24
3	Sedih	84.06
4	Marah	63.41
Rata-rata		76.1

TABEL 7. AKURASI DARI MODEL CNN YANG DIBUAT OLEH SURAJ TRIPATHI

Untuk Rata-rata dari tiap model klasifikasi yang dibuat bisa dilihat pada Tabel 8.

No	Tahun	Model	Dataset	Akurasi (rata rata per emosi)
1	2017	SVM: LIBSVM	EmoDB	79.22%
2	2017	SVM : LIBSVM	RED	50.38%
3	2017	SVM	TEDLIUM v2	63.0%
5	2017	CNN	TEDLIUM v2	66.1%
6	2018	SVM: LIBSVM	Linguistic Data Consortium	79.22%
7	2018	SVM : LIBSVM	UGA	50.38%
8	2019	CNN	IEMOCAP	76.1%
9	2019	CNN 1D	EmoDB	92%
10	2019	CNN 1D	IEMOCAP	79.72%
11	2019	CNN 2D	EmoDB	95.33%
12	2019	CNN 2D	IEMOCAP	85.58%

TABEL 8. RATA-RATA AKURASI DARI TIAP MODEL KLASIFIKASI

E. Perbandingan

Berdasarkan hasil dari *database* yang telah digunakan, akurasi yang tinggi tidak terpengaruh dari jumlah data yang telah digunakan di dalam *database*, dikarenakan ada banyak faktor lainnya yaitu, kejernihan sebuah suara dan pembagian dataset yang merata [4]. Dario Bertero melakukan percobaan menggunakan data yang cukup banyak dengan jumlah total

lebih dari 5000 data, namun akurasi yang dihasilkan belum cukup akurat. Akurasi yang diterima sebesar 66.1%. Berdasarkan persentase hasil yang belum cukup akurat dikarenakan ada salah satu emosi yang terlalu banyak datanya yaitu emosi senang, karena data yang diterima terlalu banyak dicurigai bahwa emosi "senang" berisi emosi "netral" dikarenakan saat suara yang dikeluarkan menyerupakan emosi senang dan netral *database* tersebut memiliki frekuensi yang cukup signifikan [4].

Nancy Menwal [6] mengatakan walaupun menggunakan 535 file saja, dapat menghasilkan akurasi yang lebih tinggi yaitu 80%. Hal ini dikarenakan pembagian file emosi pada *database* tersebut bisa dibilang cukup rata [6].

Jianfeng Zhao [2] membuktikan ketika menggunakan *database* yang sama dengan Nancy Memwal yaitu EmoDB, kemudian mendapatkan akurasi dengan persentase 95.33% [2].

Jika dilihat dari metode fitur ekstraksi yang digunakan, MFCCs memang sering sekali digunakan [1-2]. Manas Jain [1] salah satu orang yang menggunakan MFCCs dalam percobaan beliau, dan mendapatkan akurasi hingga 98% di salah satu emosinya, beliau juga membandingkan jika menggunakan LPCC dalam menggunakan metode ekstraksinya, dan akurasinya menurun lumayan jauh, hingga di angka 70% [1]. Tetapi, dalam percobaan Udit Jain [3], ketika beliau membandingkan LPCC dan MFCCs, LPCC memiliki akurasi yang lebih baik, yaitu di 28%, sedangkan MFCCs mendapatkan 24% [3]. Lalu beliau mencoba menggabungkan kedua metode tersebut dan membuat akurasinya pun melambung tinggi di 56% [3]. Dari pembicaraan diatas membuktikan bahwa MFCCs tidak selalu lebih baik dari LPCC.

Dilihat dari metode fitur ekstraksi yang digunakan (MFCCs) dominan digunakan, lalu Manas Jain [1] merupakan salah satu orang yang menggunakan metode MFCCs. Manas Jain melakukan percobaan sehingga mendapatkan akurasi pada salah satu emosi dengan persentase 98.52%, Manas Jain [1] membandingkan metode MFCCs dengan LPCC dalam menggunakan metode ekstraksi nya, kemudian akurasi yang dihasilkan menurun, hingga menyentuh pada persentase 70%. Namun, Udit Jain melakukan percobaan dan membandingkan antara LPCC dan MFCCs kemudian LPCC memiliki akurasi yang lebih baik, dengan persentase 28%, dibandingkan MFCCs hanya mendapatkan persentase diangka 24% [3]. Udit Jain mencoba menggabungkan kedua metode dan mendapatkan akurasi yang melambung tinggi dengan persentase diangka 56% [3]. Implikasi perbandingan diatas dari kedua metode antara MFCCs dan LPCC membuktikan bahwa MFCCs tidak selalu lebih baik dari LPCC.

Berdasarkan dari total literatur review yang telah saya kaji. Jianfeng Zhao [2] dan Nancy Semwal [6] menggunakan *database* yang sama, Jianfeng Zhao menggunakan model CNN dan Nancy Remwal menggunakan model SVM. Akurasi yang dimiliki oleh CNN, rata rata akurasi yang dihasilkan lebih tinggi dengan persentase 95.33%. Berikut untuk detail akurasi CNN dapat dilihat pada Tabel 7.

No	Emosi	Akurasi (%)
1	Marah	100
2	Bosan	97.53
3	Jijik	86.96
4	Takut	97.1
5	Senang	91.55
6	Netral	93.67
7	Sedih	98.39
Rata-rata		95.33

TABEL 9. AKURASI DARI MODEL CNN OLEH JIANFENG ZHAO MENGGUNAKAN EMODB SEBAGAI DATABASE

Pada model SVM akurasi yang didapatkan, rata-rata menghasilkan akurasi dengan persentase 80%. Untuk detail akurasi SVM dapat dilihat pada Tabel 8.

No	Emosi	Akurasi (%)
1	Marah	84.2
2	Bosan	82.7
3	Gelisah	85.5
4	Senang	67.6
5	Netral	73.41
6	Sedih	93.54
Rata-rata		80

TABEL 10. AKURASI DARI MODEL SVM OLEH NANCY SEMWAL MENGGUNAKAN EMODB SEBAGAI DATABASE

IV. KESIMPULAN

Berdasarkan data yang telah dikumpulkan. Perkembangan dalam mendeteksi sebuah sinyal suara berdasarkan emosi berkembang cukup pesat. Pada penelitian Jiangfezao dalam mendeteksi sebuah sinyal suara dapat mencapai akurasi yang cukup tinggi, dengan persentase 95.33%. Dalam pendeteksian emosi berdasarkan suara ini, sangat berpengaruh berdasarkan dengan memilih *database* yang benar. Dikarerena kan terbukti dari beberapa penelitian terdahulu, dengan menggunakan *database* yang kurang bagus, maka sangat mempengaruhi hasil dari akurasi pada setiap model yang digunakan. Untuk penggunakan model *Support Vector Machine* (SVM) dan (CNN) sudah cukup bisa dikatakan layak untuk digunakan sebagai model yang tepat untuk medeteksi emosi melalui suara.

V. REFERENSI

[1] Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, dan Rajesh Kumar Muthu., Speech Emotion Recognition using *Support Vector Machine*, International Journal of Smart Home (2017)

[2] Jianfeng Zhao, Xia Mao, Lijiang Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control, Elsevier Ltd (2019)

[3] Udit Jain , Karan Nathani , Nersisson Ruban , Alex Noel Joseph Raj, Zhemin Zhuang, dan Vijayalakshmi G V Mahesh , Cubic SVM

Classifier Based Feature Extraction and Emotion Detection from Speech, Proceedings - 2018 International Conference on Sensor Networks and Signal Processing, SNSP (2018)

[4] Dario Bertero, Pascale Fung., A first look into a *Convolutional Neural Network* for speech emotion detection , ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings (2017)

[5] Nancy Semwal, Abhijeet Kumar, dan Sakthivel Narayanan., Automatic Speech Emotion Detection System using Multi-domain Acoustic Feature Selection and Classification Models , 2017 IEEE International Conference on Identity, Security and Behavior Analysis, ISBA (2017)

[6] Suraj Tripathi , Abhay Kumar , Abhiram Ramesh , Chirag Singh , dan Promod Yenigalla., Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions, (2019)