# Ulasan : Pengenalan Emosi Melalui Suara

*by* Rio Galang

**Submission date:** 24-Nov-2020 01:10PM (UTC+0700)
**Submission ID:** 1455088184
**File name:** Kajian_Pustaka_-_17523118.pdf (166.24K)
**Word count:** 3606
**Character count:** 21723

# *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 sudah cukup banyak dilakukan oleh beberapa peniliti, seperti mendeteksi suara hewan, ataupun objek tertentu. Namun, sekarang peneliti sedang gencar sekali dalam meneliti tentang emosi manusia, dan salah satunya adalah mendeteksi emosi manusia melalui suara. Pendeteksian emosi saat ini tidak hanya digunakan untuk riset di bidang akademik seperti, psikologi, neuroscience, psikiater, ilmu kognitif dan lainya. Tetapi ditemukan juga pengaplikasian praktis seperti, call centre, gaming industry, bidang medis dan lainya Literatur ini akan menjelaskan tentang perkembangan pendeteksian emosi melalui suara dari tahun 2017 hingga 2019. Perkembangan pendeteksian emosi melalui suara sudah cukup baik, bahkan cukup banyak yang mencapai akurasi hingga 90% lebih. Model SVM maupun CNN sudah cukup layak digunakan sebagai model yang digunakan untuk mendeteksi emosi melalui suara***

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

## PENDAHULUAN

Penggunaan suara dalam *data science* sudah cukup banyak dilakukan oleh beberapa peneliti, seperti mendeteksi suara hewan, ataupun objek tertentu. Namun, sekarang peneliti sedang gencar sekali dalam meneliti tentang emosi manusia, dan salah satunya adalah mendeteksi emosi manusia melalui suara[1]. Pendeteksian emosi saat ini tidak hanya digunakan untuk riset di bidang akademik seperti, psikologi, *neuroscience*, psikiater, ilmu kognitif dan lainya. Tetapi ditemukan juga pengaplikasian 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 gimik 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]. Bisa terjadi suatu kemungkinan seperti ini, dia 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, Kajian pustaka di sini 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.

## SUMBER METODE

a. Portal Literatur

Literatur diperoleh melalui beberapa portal edukasi yang sudah terpercaya dan mudah dilacak kembali seperti Google Scholar. tersebut dibatasi dalam tahun 2017-2020. Fokus pencarian literatur menggunakan *keyword* emotion, speech, dan recognition.

b. Batasan dalam Seleksi Literatur

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.

## HASIL DAN DISKUSI

*a*. *Database* Suara

Pada tahun 2017, Dario Bertero menggunakan dataset yang berasal dari TEDLIUM v2 corpus. Di dalam *database* tersebut terdapat 5000 lebih data yang berisikan beberapa emosi yaitu, sedih, marah, senang, netral, dan garbage. Data dari database tersebut sebagian besarnya disuarakan oleh murid dari research group yang Dario Bertero miliki (sekitar 90%), sisanya adalah data crowdsourcing dari Amazon Mechanical Turk . Dari 5000 lebih data tersebut, pembagiannya adalah sebagai berikut, 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]. Pada tahun yang sama, Nancy Semwal [5] menggunakan dataset yang 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. Tetapi, dalam penelitian beliau, emosi jijik dihapus sehingga tersisakan 489 data. RED

sendiri berisikan 720 data dari enam emosi yang berbeda, dan setiap emosinya terbagi rata semua di angka 120 data. Emosi yang tersedia di dalam database tersebut adalah marah, jijik, takut, senang, sedih, dan terkejut. Berbeda dari database EmoDB, database ini berisikan berbagai bahasa, yaitu inggris, mandarin, urdu. persian, italian dan punjabi.

Berlanjut ke 2018, Manas Jain [1] menggunakan dua database yang berbeda pula, yaitu Linguistic Data Consortium (LDC), dan UGA. Kedua database tersebut memiliki jumlah data yang sama, yaitu di 100 data. Emosi yang tersedia di tiap database pun sama yaitu, sedih, senang, marah, dan netral[1]. Udit Jain [3] juga melakukan penelitian di tahun 2018, beliau mengambil dataset yang berasal dari platform yang cukup besar yaitu Youtube. Dari Youtube tersebut, beliau mengunduh rekaman suara dari 20 artis yang berbeda, 10 laki laki dan 10 perempuan. Suara yang diunggah tersebut berisi artis yang mengucapkan beberapa emosi. Emosi yang diambil dari Youtube ada empat, yaitu senang, sedih, marah, dan netral. Di tiap emosi rekaman tersebut berdurasi dua detik dan berformat .wav. Jumlah data yang didapatkan ada 400 audio file, 200 suara laki laki dan 200 lainnya suara perempuan. Pembagian per emosi untuk database tersebut bisa dibilang cukup rata, tiap emosi terdapat 50 file [3].

Kemudian di tahun 2019, Jianfeng Zhao [2] menggunakan dua *database* yang berbeda pula, yaitu EmoDB (database yang sama saat penelitian Nancy Semwal) dan Interactive Emotional Speech Dataset (IEMOCAP). Untuk *database* EmoDB masih sama saat diteliti oleh Nancy Semwal di tahun 2017. Seperti database tersebut memiliki 535 *files* dan tujuh emosi yang berbeda. IEMOCAP memiliki data sebanyak 1150, data tersebut disuarakan oleh lima penyuara (laki laki dan perempuan).Emosi yang tersedia dari database tersebut ada tujuh, yaitu marah, excited, frustasi, senang, netral, sedih, dan terkejut. Detail pembagian suara dari database adalah sebagai berikut, 71 suara masuk ke dalam kelas “marah”, 178 suara masuk ke kelas “excited”, 271 suara masuk ke kelas “frustasi”, 31 suara masuk ke kelas “senang”, 328 suara masuk ke kelas “netral, 265 suara masuk ke kelas “sedih”, dan 6 suara masuk ke kelas “terkejut”.

Suraj Tripathi [6] juga melakukan penelitian di tahun 2019, beliau mengambil dataset berasal dari University of Southern California’s Interactive Emotional Motion Capture (USC-IEMOCAP). Isi dari *database* tersebut adalah percakapan antara dua orang, laki laki dan perempuan. Topik dari percakapan tersebut sudah diatur sedemikian rupa hingga menghasilkan beberapa emosi yaitu marah, senang, netral, dan sedih. Percakapan yang panjang tersebut kemudian di potong hingga menjadi 3 - 15 detik untuk mengutarakan satu emosi. Detail pembagian suara dari database tersebut adalah 48.8% suara masuk ke kelas emosi “netral”, 12.3% suara masuk ke kelas emosi “senang”, 26.9% suara masuk ke kelas emosi “sedih”, 12% suara masuk ke kelas emosi “marah” [6].

*b*. Ekstraksi Fitur

Dorio Bertero [4] dan Jianfeng Zhao [2] tidak menggunakan metode tambahan saat melakukan ekstraksi fitur, beliau mempercayakan ekstraksi fitur pada layer konvolusi. Menurut Nancy Semwal [5] menggunakan metode Low Level Descriptors (LLDs), beliau mempercayakan pada metode tersebut karena menurut beliau, metode tersebut cocok untuk mengekstraksi dari sinyal suara pada percakapan yang pendek. Menurut Manas Jain [1] menggunakan dua metode untuk mengekstraksi fitur yaitu, *Mel-Frequency cepstral coefficients* (MFCCs) dan *Linear Prediction Cepstral Coefficient* (LPCC) [1]. MFCCs digunakan karena metode tersebut sangat umum digunakan saat membahas ekstraksi fitur. Perhitungan metode tersebut berdasarkan dari karakter telinga manusia, dan ini menyebabkan perhitungannya mirip dengan sistem pendengaran manusia. LPCC digunakan untuk mengulas dari setiap *envelope* yang berisi sinyal suara percakapan dalam bentuk 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 intensitan dan frekuensi dari sinyal yang tersisa [1]. Suraj Tripathi [6] tidak hanya bergantung pada layer konvolusi saat ekstraksi fitur, beliau mencoba menggunakan MFCCs. Pada saat ingin menggunakan MFCCs beliau menggunakan phyton *package* bernama librosa [6]. Udit Jain [3] tidak hanya menggunakan satu metode ekstraksi fitur, namun menggunakan hingga tiga ekstra fitur. Pertama adalah MFCCs, metode ini memberikan akurasi yang tinggi saat pengklasifikasian karena sistem ini sangat mempertimbangkan sensitivitas persepsi manusia terhadap frekuensi dari pengenalan itu. Kedua adalah LPCC, prinsip penghitungan koefisien melalui metode ini adalah setiap sampel suara pada sumbu waktu x saat ini dapat diproyeksikan sebagai kombinasi linier dari sampel suara sebelumnya . Di sini saluran vokal dimodelkan menggunakan filter digital all-pole. Ketiga adalah Formant Frequency (FF), Konsentrasi energi akustik dalam rentang frekuensi tertentu dapat didefinisikan sebagai forman gelombang wicara tertentu dan mewakili kekuatan sinyal wicara yang diperoleh pada frekuensi tertentu. Pendekatan khusus fungsi penundaan akar spektral diimplementasikan untuk ekstraksi forman di mana operasi () r digunakan sebagai pengganti fungsi logaritmik [3].

*c*. Klasifikasi

Dario Bertero [4] menggunakan Convolutional Neural Network (CNN) sebagai model pengklasifikasian. Beliau menggunakan hanya satu lapisan pada saat konvolusi, hal tersebut membuat waktu evaluasi menjadi lebih cepat beberapa mili sekon dibandingkan menggunakan struktur dua lapisan. Lapisan konvolusi tersebut bertugas untuk mengekstraksi fitur di tiap framenya dan mengevaluasi perbedaan dari frame yang *overlapping*. Setelah konvolusi terdapat proses *max-pooling*. *Max-pooling* mengizinkan kita untuk memilih kontribusi dari frame yang paling signifikan, dan untuk mengkombinasi semuanya menjadi vector yang sudah tetap ukurannya. Selanjutnya diikuti lapisan fully connected (FC) (sebanyak 200) dan terakhir ada lapisan softmax yang melakukan klasifikasi yang sebenarnya. Tetap

bagaimanapun *deep learning* berbasis metode yang bisa disebut "*magical boxes*" tetapi tetap menghasilkan hasil yang bagus. Saat deep learning digunakan untuk mempelajari representasi fitur selain hanya melakukan klasifikasi, lebih penting lagi untuk memverifikasi apa yang terjadi di dalamnya. Beliau sempat membandingkan akurasi yang dimiliki jika menggunakan SVM dan CNN, hasilnya terlihat di rata rata akurasinya CNN lebih unggul sedikit. [4]

Nancy Semwal [5] menggunakan SVM sebagai model pengklasifikasiannya. SVM dapat mentransfer sekumpulan fitur yang ada dalam vektor, setiap vektor akan diberikan tanda untuk masuk ke satu kategori atau dua kategori, ke dalam high dimensional space sehingga vektor yang sekarang dipindahkan ke tempat yang baru, dan sudah dipisahkan dari setiap kategori sejauh mungkin. Proses tersebut dilakukan dengan menggunakan *kernel functions*. SVM sangat efektif di dalam kasus fitur dimensi lebih tinggi daripada jumlah sampel, membuat SVM menjadi sangat cocok digunakan. Pada penelitian kali ini beliau menggunakan implementasi dari penggunaan library bernama LIBSVM [5].

Manas Jain [1] menggunakan SVM sebagai model pengklasifikasian. Beliau berpendapat bahwa model SVM ini sangat simpel dan efisien dalam algoritma pengklasifikasian ketika digunakan untuk mengklasifikasi dan pengenalan pola. Di dalam SVM sendiri memiliki dua jenis yaitu, linear dan tidak linear. Beliau menggunakan LIBSVM sebagai *library* yang digunakan saat pengklasifikasian menggunakan SVM. Menurut beliau, saat model sudah selesai disiapkan, akan menjadi sangat mudah untuk memprediksi emosi dari testing dataset [1].

Udit Jain[3] menggunakan SVM sebagai model pengklasifikasian. Sebelum masuk ke model SVM, beliau menggunakan *k-fold Cross-Validation* untuk mensegmentasi data tersebut untuk masuk ke training database dan testing database. Setelah melalui segmentasi tersebut baru data tersebut dimasukkan ke dalam model SVM. Berbeda dari SVM yang biasa, beliau menggunakan Cubic SVM sebagai model pengklasifikasian. Cubic SVM dipilih karena model tersebut memiliki waktu komputasi lebih sedikit 10 detik dari model SVM pada umumnya [3].

Suraj Tripathi [6] juga menggunakan model CNN, namun pada layer konvolusinya terdapat empat paralel. Pada saat konvolusi beliau menggunakan 200 kernel di tiap parallel. Untuk mencegah memilih kernel yang ukurannya bisa jadi kurang optimal, maka beliau memutuskan untuk membuat kernel yang ukurannya berbeda beda. Selanjutnya setelah dilakukan konvolusi di layer pertama, maka dikirim ke max pool layer, setelah itu dikirim ke Fully lapisan Connected (FC), setelah itu dikirim ke layer terakhir yaitu *Softmax Layer* disinilah proses pembagian empat emosi yang berbeda [6]. Jianfeng Zhao [2] menggunakan dua macam CNN sebagai model pengklasifikasian. Pertama adalah 1D CNN Long Short - Term Memory (LSTM) ,1D CNN LSTM ini disusun dengan menumpuk empat local feature learning blocks (LFLBs), satu lapisan LSTM dan satu lapisan FC. Model ini didesain untuk mempelajari fitur yang dalam dari audio yang bersifat raw. Maka dari itu, konvolusi dan *pooling kernel* di tiap LFLB hanya satu dimensi. Konvolusi kernel di tiap LFLB memiliki ukuran yang sama yaitu tiga, dan punya langkah yang sama yaitu satu, dan memiliki lapisan yang sama. Ketika suara, direpresentasikan sebagai satu dimensional vektor, maka suara tersebut akan dipindah ke model ini, kemudian fitur dari suara tersebut akan dipelajari oleh LFLBs. Setelah diubah bentuknya, fitur tersebut akan dimasukkan ke lapisan LSTM. Setelah itu fitur yang telah dipelajari tersebut akan dipindahkan ke lapisan FC. Kedua adalah 2D CNN LSTM, 2D CNN LSTM mempunyai struktur yang sama dengan 1D CNN LSTM. Perbedaannya adalah konvolusi dan pooling kernel di tiap LFLB adalah dua dimensi. Jumlah konvolusi kernel di LFLBs pertama dan kedua adalah 64, lalu di ketiga dan keempat adalah 128. Model ini didesain untuk mempelajari high-level fitur dari emosi yang berasal dari log-mel spektogram. Ketika log-mel spektogram dalam bentuk matriks dimasukkan ke dalam model ini, fitur lokal dengan korelasi lokal akan dipelajari oleh empat LFLBs. Fitur yang dikeluarkan dari LFLBs akan dibentuk ulang menjadi sekuensi yang sementara dan dimasukkan ke lapisan LSTM . Kemudian dependensi kontekstual akan belajar dari lokal fitur. Lapisan FC akan menggeneralisasi dari fitur fitur yang telah dikeluarkan, dan Softmax akan diadopsi untuk membuat prediksi dari fitur yang dipelajari tadi[2].

*d*. Akurasi

Pada tahun 2017, 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].

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].

Kemudian di tahun 2018, Manas Jain [1] mendapatkan akurasi rata rata 85.085% ketika menggunakan MFCCs sebagai metode dalam ekstraksi fitur, lalu mendapatkan akurasi rata rata 73.125% ketika menggunakan LPCC sebagai metode dalam ekstraksi fitur. Beliau juga membandingkan jika menggunakan metode MFCCs lalu dengan dua *database* yang berbeda, hasilnya sebagai berikut, saat menggunakan database dari LDC, beliau mendapatkan detail akurasi 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]. 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]. Berlanjut ke tahun 2019, 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]. Masih di tahun 2019, 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].

### *A. Perbandingan*

Jika dilihat dari database yang digunakan, jumlah tidak selalu mempengaruhi hasil dari akurasi yang tinggi, ada banyak faktor lainnya, yaitu kejelasan suara, dan pembagian dataset yang merata [4]. Contohnya seperti percobaan yang dilakukan oleh Dario Bertero [4], beliau menggunakan data yang cukup banyak lebih dari 5000 data, namun akurasi yang dihasilkan tidak cukup akurat hanya di 66.1%. Hal tersebut dikarenakan ada salah satu emosi yang terlalu banyak datanya yaitu emosi senang, karena hal tersebut dicurigai bahwa emosi "senang" tersebut ada berisi emosi "netral", dikarenakan saat yang menyuarakan emosi "senang" dan "netral" dalam database tersebut memiliki frekuensi yang cukup sama [4]. Nancy Menwal walaupun menggunakan 535 file saja, bisa menghasilkan akurasi yang lebih tinggi di 80 %. Hal ini dikarenakan pembagian file emosi pada database tersebut bisa dibilang cukup rata. [6]. Ini dibuktikan kembali ketika Jianfeng Zhao [2] menggunakan database yang sama dengan Nancy Menwal yaitu EmoDB, mendapatkan akurasi hingga 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 99%, 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.

Sekian dari 6 makalah yang saya baca, hanya Jianfeng Zhao [2] dan Nancy Remwal [5] yang menggunakan database yang sama. Jianfeng Zhao menggunakan model CNN dan Nancy Remwal menggunakan model SVM. CNN ternyata memiliki rata rata akurasi yang lebih tinggi di angka 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" [2]. Sedangkan jika menggunakan SVM, bisa mendapatkan rata rata akurasi hanya di 80%. Untuk detail akurasi sebagai berikut , 84.2% ketika emosi "marah", 85.5% ketika emosi "gelisah",82.7% ketika emosi "bosan", 67.6% ketika senang, 73.41% saat "netral", dan 93.54% file saat "sedih" [5]. Dilihat dari data tersebut CNN selalu memiliki akurasi yang lebih tinggi dibandingkan SVM.

TABEL 1. PERBANDINGAN TIAP MODEL UNTUK MENDETEKSI EMOSI MELALUI SUARA

| 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% |

## KESIMPULAN

Dari data data diatas dapat disimpulkan bahwa, perkembangan pendeteksian emosi melalui suara sudah cukup baik, bahkan cukup banyak yang mencapai akurasi hingga 90% lebih. Hal yang cukup berpengaruh dalam pendeteksian emosi adalah memilih database yang benar. Hal tersebut terbukti dari percobaan Dorio Bertero [4], karena beliau memilih database yang kurang bagus, maka membuat akurasi dari model yang ia buat kurang baik. Model SVM maupun CNN sudah cukup layak digunakan sebagai model yang digunakan untuk mendeteksi emosi melalui suara.

## 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)

# Ulasan : Pengenalan Emosi Melalui Suara

ORIGINALITY REPORT

| 6% | 5% | 5% | 0% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | dblp.dagstuhl.de<br>Internet Source | 1% |
| 2 | export.arxiv.org<br>Internet Source | 1% |
| 3 | Alifia Revan Prananda, Hanung Adi Nugroho, Igi Ardiyanto. "Enumeration of Plasmodium Parasites on Thin Blood Smear Digital Microscopic Images", 2019 5th International Conference on Science in Information Technology (ICSITech), 2019<br>Publication | 1% |
| 4 | dblp.uni-trier.de<br>Internet Source | 1% |
| 5 | esajournals.onlinelibrary.wiley.com<br>Internet Source | 1% |
| 6 | Udit Jain, Karan Nathani, Nersisson Ruban, Alex Noel Joseph Raj, Zhemin Zhuang, Vijayalakshmi G.V. Mahesh. "Cubic SVM Classifier Based Feature Extraction and | <1% |

Emotion Detection from Speech Signals", 2018 International Conference on Sensor Networks and Signal Processing (SNSP), 2018
Publication

---

7 "Advances in Computational Intelligence Techniques", Springer Science and Business Media LLC, 2020 <1%
Publication

---

8 Nancy Semwal, Abhijeet Kumar, 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 <1%
Publication

---

9 Cahyo Dwi Raharjo, Izzati Muhimmah. "Remodeling of human foot using chain code for designing special shoes", 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2015 <1%
Publication

---

10 Zbancioc Marius Dan, Feraru Silvia Monica. "A study about MFCC relevance in emotion classification for SRoL database", 2013 4th International Symposium on Electrical and Electronics Engineering (ISEEE), 2013 <1%
Publication

---

| | | |
|---|---|---|
| 11 | menyusun-instrumen-riset-kuantitatif.blogspot.com<br>Internet Source | <1% |
| 12 | artikelterbaru.com<br>Internet Source | <1% |
| 13 | informatics.uii.ac.id<br>Internet Source | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |