Sistem Pengenalan Suara untuk Perintah pada Robot Sepak Bola Beroda

Muhammad Azhar Ismail, Dhiaul Ma'ruf, Muhammad Revo Khairullah, Fajar Budiman, Rudy Dikairono, Djoko Purwanto

Institut Teknologi Sepuluh Nopember (ITS)

Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia

djoko@ee.its.ac.id

Abstrak— Pemberian perintah kepada robot sepak bola beroda merupakan salah satu hal penting yang dapat digunakan untuk menentukan strategi permainan pada robot. Penyampaian perintah dari manusia menuju robot disebut dengan *high level human coaching*. Salah satu metode yang dapat digunakan untuk melakukan hal tersebut adalah dengan memberikan perintah suara. Pada penelitian tugas akhir dirancang sebuah sistem yang dapat digunakan untuk mengenali perintah suara yang diberikan oleh manusia kepada robot. Mikrofon digunakan sebagai sensor untuk menangkap perintah suara yang disampaikan. Selain itu digunakan metode VAD (*Voice Activity Detector*) untuk merekam perintah suara yang disampaikan. Fitur pada setiap rekaman suara akan diambil menggunakan metode MFCC *(Mel Frequency Cepstral Coefficients*). Fitur tersebut akan digunakan sebagai input dari CNN (*Convolutional Neural Network*). Metode CNN digunakan agar sistem dapat mengenali setiap perintah suara yang disampaiakn berdasarkan fitur yang didapatkan dari proses MFCC. Sistem pengenalan suara ini akan menjadikan robot dapat melakukan beberapa perintah, berdasarkan suara yang disampaikan. Berdasarkan pengujian yang telah dilakukan pada 9 orang dengan 7 suara perintah berbeda, diapatkan hasil bahwa sistem pengenalan suara ini mampu mengenali setiap perintah tersebut dengan akurasi keberhasilan rata-rata 83% untuk semua pengujian. Sedangkan untuk mengenali orang yang memberikan perintah didapat akurasi keberhasilan rata-rata sebesar 80% pada 9 orang yang dilakukan pengujian.

Kata kunci: pengenalan perintah suara, *voice activity detector* (VAD), *mel frequency cepstral coefficients* (MFCC), *convolutional neural network* (CNN)

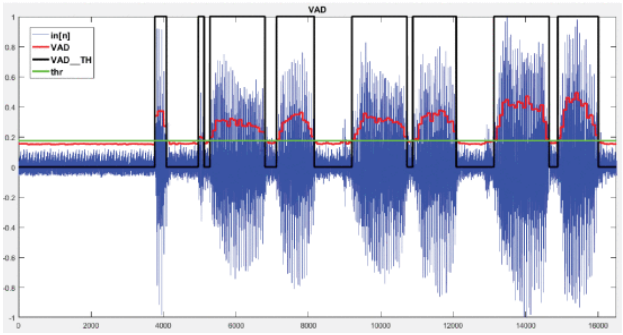
I. PENDAHULUAN

Dalam pertandingan robot sepak bola beroda, tim pemenang adalah dengan menghitung jumlah gol terbanyak yang dicetak oleh setiap tim, untuk itu dibutuhkan strategi permainan yang efektif agar dapat mencetak banyak gol. Setiap tim yang bertanding pasti memiliki strategi permainan tertentu yang berbeda antara tim satu dengan lainya. Untuk mengatasi hal tersebut, maka dibutuhkan sebuah perintah dari pelatih untuk mengubah startegi atau pola permainan ketika melawan tim yang berbeda. Perintah ini dapat disampaikan menuju robot melalui menggunakan metode pengenalan suara. Untuk melakukan pengenalan suara pada pekatih, maka dibutuhkan sebuah sistem yang dapat mengenali suara perintah tersebut. Dengan adanya sistem tersebut nantinya robot akan dapat mengenali berbagai jenis perintah yang disampaikan melalui suara. Jadi perintah yang disampaikan nantinya tidak hanya tentang strategi permainan, namun perintah tersebut juga dapat digunakan untuk memposisikan ulang robot yang bermasalah. Sistem pengenalan suara yang dibuat ini menggunakan mikrofon untuk menangkap sinyal suara disekitarnya. Kemudian terdapat sistem untuk merekam suara dengan metode *voice activity detector* (VAD) untuk membedakan suara perintah dengan *noise*. Selanjutnya fitur suara yang terekam diambil dengan menggunakan metode *mel spectrum cepstral coefficients* (MFCC). Hasil fitur yang didapat akan digunakan sebagai masukan dari *convolutional neural network* (CNN). CNN akan bertugas untuk melakukan prediksi perintah yang diterima berdasarkan data fitur dari rekaman suara yang didapat.

II. TEORI PENUNJANG

A. *Voice Activity Detector (VAD)*

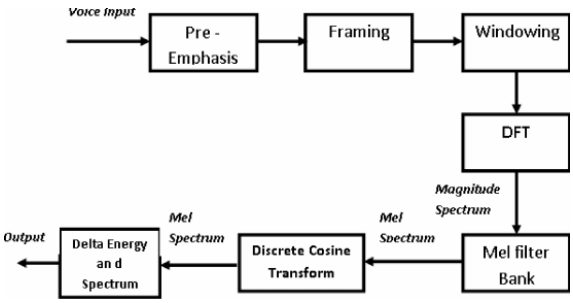
Voice activity detector merupakan sebuah metode yang dapat digunakan untuk membedakan sinyal suara dan sinyal *noise*. VAD yang baik adalah VAD yang mampu mendeteksi keberadaan suara manusia pada sebuah sinyal yang memiliki latar belakang *noise* yang cukup kuat [1]. Suatu VAD dapat dikategorikan menjadi tiga jenis yaitu *supervised learning*, *semi-supervised learning*, dan *unsupervide learning* [1]. Sistem VAD dengan sistem *learning* merupakan sistem VAD yang lebih efektif untuk mendeteksi suara manusia dengan latar belakang *noise* yang kuta. Selain itu juga terdapat metode VAD tanpa menggunakan *learning*, yaitu menggunakan nilai ambang batas rata-rata dari amplitudo sinyal yang tertangkap. Jadi suara yang akan direkam merupakan suara yang dapat melewati nilai ambang batas yang ditentukan seperti pada Gambar 1.



Gambar 1. VAD dengan Nilai Ambang Batas [2]

B. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients telah banyak digunakan dalam voice recognition. MFCC dapat digunakan sebagai ekstraksi fitur yang baik untuk mewakili ucapan manusia atau sinyal music dan yang paling penting adalah MFCC telah terbukti bermanfaat untuk pengenalan suara. Perhitungan yang dilakukan dalam MFCC menggunakan dasar dari perhitungan short-term analysis [3]. MFCC sebenarnrnya merupakan adaptasi dari sistem pendengaran manusia dimana untuk mendapatkan fitur dari sebuah sinyal suara, sinyal suara akan difilter secara linier untuk frekuensi rendah dan secara logaritmik untuk frekuensi tinggi.



Gambar 2. Proses Perhitungan MFCC **[4]**

1. Pre-emphasis

Proses *Pre-emphasis* adalah proses memindahkan sinyal yang masuk menuju sebuah filter untuk menekan atau menguatkan frekuensi tinggi pada sinyal. Proses ini akan meningkatkan energi dari sinyal pada frekuensi tinggi [4]. Filter *pre-emphasis* dapat diterapkan pada sebuah sinyal dengan filter orde satu dengan persamaan seabagai berikut:

$$y(t) = x(t) - \alpha x(t-1) \quad (1)$$

(α) adalah koefisien dari filter, nilai yang sering digunakan untuk koefisien tersebut adalah 0.95 atau 0.97.

2. Framing

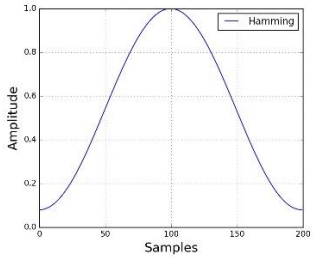
Proses segmentasi sinyal diperoleh dari konversi data analog menjadi digital (ADC) [4]. Kemudian sinyal digital tersebut dipecah menjadi beberapa frame kecil dengan Panjang 20 hingga 40 ms [5]. Alasan dilakukanya *framing* adalah karena frekuensi sinyal berubah setiap waktu, sehingga dalam banyak kasus kurang efektif bila melakukan transformasi fourier pada seluruh sinyal karena akan menghilangkan kontur frekuensi sinyal pada setiap waktunya. Untuk menghindarinya, dapat diasumsikan bahwa frekuensi suatu sinyal akan tetap pada periode waktu yang sangat singkat.

3. Windowing

Pada proses *windowing* ini fungsi yang umum digunakan adalah *hamming window*. *Hamming window* digunakan untuk memetakan sinyal pada setiap frame yang diperoleh dengan mempertimbangkan blok berikutnya pada proses ekstraksi fitur dan mengintegrasikan semua frekuensi terdekat [4]. Persamaan *Hamming window* adalah sebagai berikut:

$$w[n] = 0.54 - 0.46cos(\frac{2\pi n}{N-1}) \quad (2)$$

$0 \leq n \leq N-1$, diaman N adalah Panjang *window*. Dengan melakukan plotting pada persamaan diatas, maka akan diperoleh sebuah grafik yang dapat dilihat pada Gambar 3.



Gambar 3. *Hamming Window*

Terdapat beberapa alasan dilakukanya Hamming window pada setiap frame, alasan utamanya yaitu utuk menangkal asusmsi yang dibuat oleh fast fourier transform (FFT) bahwa data tidak terbatas dan untuk mengurangi nilai kebocoran spektral [4].

4. Fast Fourier Transform (FFT)

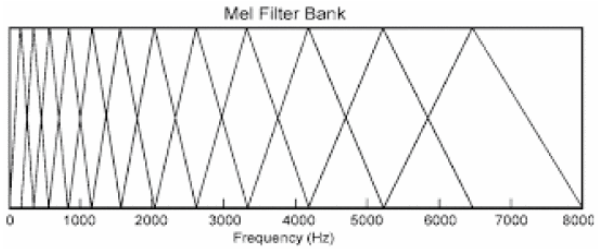
Fast fourier transform berfungsi untuk mengkonversi setiap *frame* dari domain waktu ke domain frekuensi pada seluruh sampel yang didapat. Persamaan dari FFT adalah sebagai berikut:

$$Y(w) = FFT[h(t) * x(t)] = H(w).X(w) \quad (3)$$

Y (w), H (w) dan X (w) adalah transformasi fourier dari y(t), h(t) dan x(t).

5. Mel Filter Bank

Rentang frekuensi pada spektrum FFT sangat luas dan sinyal suara tidak dapat mengikuti skala liniernya, sehingga dibutuhkan suatu filter seperti pada Gambar 4.



Gambar 4. *Mel Filter Bank* [4]

Gambar diatas menunjukkan filter segitiga yang digunakan untuk menghitung jumlah komponen spektral filter, sehingga output proses dapat mendekati skala Mel. Setiap respon frekuensi magnitudo filter berbentuk segitiga dan sama pada frekuensi tengah dan berkurang secara linear menjadi nol pada daua frekuensi tengah filter yang berdekatan [4].

6. Discrete Cosine Transform (DCT)

DCT berfungsi untuk mengkonversi spektrum log *mel* kembali dalam domain waktu. Representasi *cepstral* dari spektrum suara memberikan representasi yang bagus pada sifat spektral lokal sinyal yang dihasilkan dari analisa *frame* [4]. Karena koefisien spektrum mel adalah bilangan real, maka kita dapat mengkonversinya ke dalam domain waktu dengan *discrete cosine transform* (DCT). DCT dapat mengkonverikan log mel kembali ke waktu. Hasil dari perhitungan DCT akan

menghasilkan koefisien frekuensi spektrum *mel* (MFCC). Berikut adalah persamaan dari *discrete cosine transform*:

$$Cn = \sum_{k=1}^{k} (\log S_k)\cos(\mathrm{n}.(\mathrm{k} - \tfrac{1}{2}).\tfrac{\pi}{k} \quad (4)$$

n = 1,2,…..k, dimana S_k,k = 1,2….k adalah output dari langkah terakhir.

7. Delta Energi dan Delta Spektrum

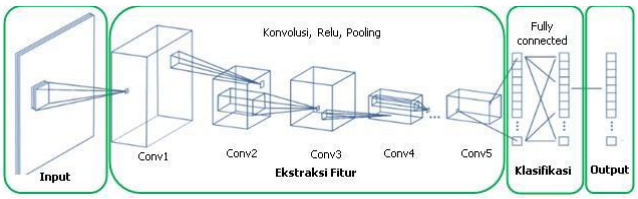
Energi merupakan sesuatu yang berhubungan dengan identitas suara dan ini bermanfaat sebagai isyarat deteksi suara [4]. Energi pada sebuah *frame* untuk sebuah sinyal x didalam *window* untuk rentang waktu t1 dan t2 dapat direpresentasikan dengan persamaan berikut:

$$Energy = \sum_{t=t1}^{t2} \mathrm{x}^2(t) \quad (5)$$

Selain itu, sinyal suara juga tidak konstan pada setiap *frame*. Ini merupakan fakta penting tentang sinyal suara pada perubahan *frame*. Untuk alasan ini dibutuhkan penambahan fitur yang berhubungan dengan pada perubahan fitur *cepstral* pada setiap waktu. Fitur yang dapat ditambahkan yaitu delta dan *double* delta [4].

C. Convolutional Neural Network (CNN)

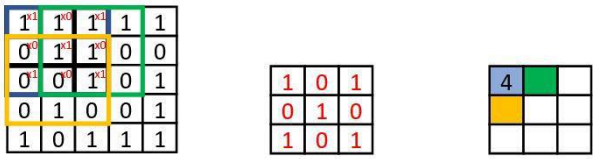
Convolutional neural network dapat disebut sebagai kelanjutan atau pengembangan dari model jaringan saraf tiruan tradisional. *Convolutional Neural Network* termasuk dalam *deep learning* karena kedalaman jaringan atau *neuron* yang tinggi [6]. *Deep learning* adalah cabang dari *machine learning* yang dapat mengajarkan komputer untuk melakukan pekerjaan selayaknya manusia, seperti komputer dapat belajar dari proses *training* [7]. CNN merupakan operasi konvolusi yang menggabungkan beberapa lapisan pemrosesan, menggunakan beberapa elemen yang beroperasi secara paralel dan terinspirasi oleh sistem saraf manusia. Arsitektur CNN dapat dilihat pada Gambar 5.



Gambar 5. Arsitektur CNN [7]

1. Convolutional Layer

Convolution Layer melakukan operasi konvolusi pada data yang dimasukan. *Layer* tersebut adalah proses utama yang mendasari sebuah CNN. Konvolusi adalah suatu istilah matematis yang berati mengaplikasikan sebuah fungsi pada output fungsi lain secara berulang. Dalam pengolahan citra, konvolusi berati mengaplikasikan sebuah filter pada data. Filter akan bergerak pada seluruh bagian data. Tujuan dilakukannya konvolusi pada data adalah untuk mengekstraksi fitur dari data input [6]. Konvolusi akan menghasilkan transformasi linear dari data input sesuai informasi pada data.

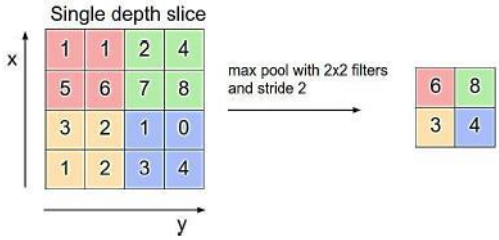


Gambar 6. Proses Konvolusi [7]

Terdapat parameter yang dapat diubah untuk memodifikasi sifat tiap lapisan, yaitu ukuran filter, *stride* dan *padding*. *Stride* mengontrol bagaimana filter diterapkan pada data input dengan bergerak sepanjang ukuran data yang telah ditentukan [7]. *Padding* adalah penambahan ukuran data dengan nilai tertentu disekitar data input agar hasil dari bidang *receptive* tidak terlalu kecil sehingga tidak banyak informasi yang hilang [7].

2. Pooling Layer

Pooling atau *subsampling* merupakan metode melakukan pengurangan ukuran matriks. *Pooling* juga bertujuan untuk meningkatkan invariansi posisi dari fitur [6]. Terdapat dua macam *pooling* yang sering digunakan yaitu *average pooling* dan *max pooling*. Namun yang paling sering digunakan yaitu *max pooling*, dapat dilihat pada Gambar6.



Gambar 7. Operasi *Max Pooling* [6]

Proses tersebut memastikan fitur yang didapatkan akan sama meskipun objek data mengalami pergeseran.

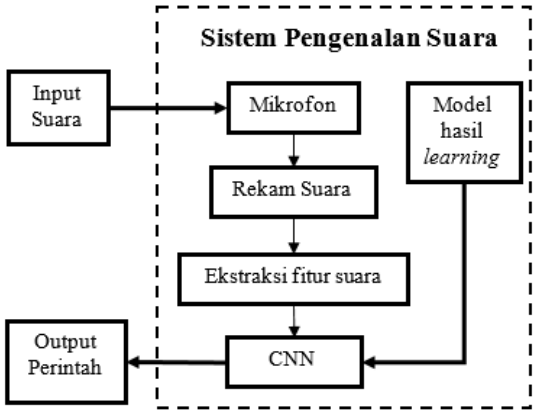
3. Fully Connected Layer

Fully connected layer merupakan kumpulan dari proses konvolusi yaitu merubah data hasil konvolusi menjadi satu dimensi. Lapisan ini mendapatkan input dari proses sebelumnya untuk menentukan fitur mana yang paling berkorelasi dengan kelas tertentu. Fungsi dari lapisan ini adalah untuk menyatukan semua *node* menjadi satu dimensi [7]. *Fully connected layer* hanya dapat diimplementasikan di akhir proses konvolusi.

III. PERANCANGAN SISTEM

Sistem pengenalan suara yang dirancang memiliki beberapa tahapan untuk mengenali sebuah suara yang masuk. Proses dari penegenalan suara dapat dilihat pada Gambar 8. Pertama suara dari manusia akan ditangkap mikrofon, kemudian suara tersebut akan direkam dan disimpan. Selanjutnya rekaman suara suara tersebut diambil fiturnya dengan metode *Mel Frequency Cepstral Coefficients* (MFCC). Data hasil dari ekstraksi fitur tersebut kemudian akan dijadikan input dari *Convolutional Neural Network* (CNN). Hasil dari proses CNN adalah berupa

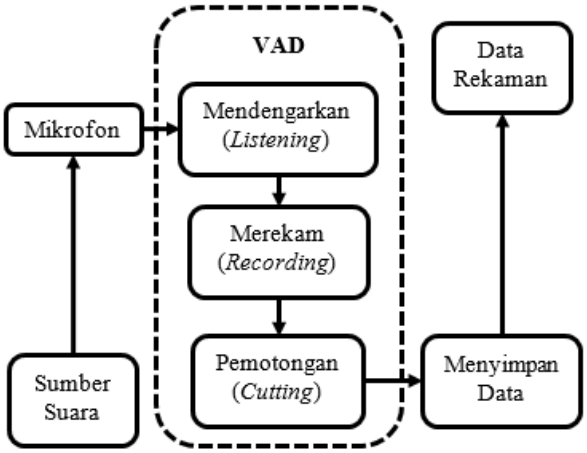
prediksi perintah yang terekam. Perintah tersebut nantinya akan disampaikan kepada robot speak bola beroda.



Gambar 8. Diagram Blok Sistem Pengenalan Suara

A. *Voice Activity Detector (VAD)*

Frekuensi *sampling* sinyal suara yang digunakan yaitu 44100 Hz. Artinya terdapat 44100 data yang diambil dari mikrofon selama satu detik. Untuk sistem rancangan VAD yang digunakan dapat dilihat pada Gambar 9.



Gambar 9. Diagram Blok Proses Merekam Suara

Agar sebuah sinyal suara dapat terekam, maka sinyal suara tersebut harus dapat melebihi nilai ambang batas yang telah ditentukan. Sistem VAD yang dibuat ini akan selalu aktif untuk menangkap sinyal suara disekitarnya, namun hanya sinyal yang dapat melebihi nilai ambang batas yang akan direkam dan disimpan. Suara yang terekam akan disimpan dalam *waveform audio format* (.WAV).

B. *Ekstraksi Fitur Suara*

Untuk mengambil fitur pada data rekaman digunakan metode *mel frequency cepstral coefficients* (MFCC). Fitur yang dihasilkan oleh MFCC adalah koefisien spektrum frekuensi *mel*. Berikut beberapa tahap untuk mengambil fitur suara dengan MFCC:

1. Koefisien yang diambil pada ekstraksi fitur suara ini berjumlah 20.
2. Waktu *sampling* yang digunakan untuk menghitung spektrum frekuensi adalah 1.5ms.
3. Hasil dari ekstraksi fitur MFCC ini dibatasi pada durasi 1.2 detik agar jumlah spektrum sama.
4. Apabila durasi rekaman kurang dari 1.2 detik, maka hasil fitur MFCC rekaman tersebut akan ditambah dengan nilai nol sampai durasinya menjadi 1.2 detik. Sedangkan untuk rekaman suara yang lebih dari 1.2 detik, maka rekaman suara tersebut akan dipotong pada durasi 1.2 detik awal.

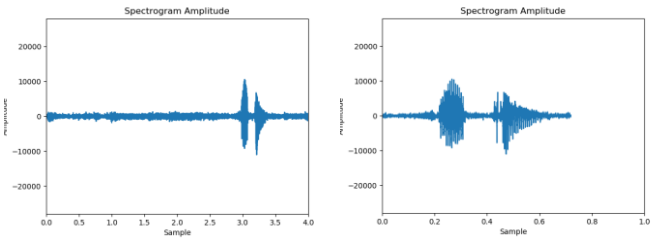
C. *Model Convolutional Neural Network (CNN)*

Model CNN yang dibuat ini, dirancang untuk mengenali perintah dan orang dari suara yang dimasukan. Data yang dilatihkan pada CNN ini berasal dari suara 5 orang laki-laki, masing-masing mengucapkan 7 perintah yang sama, yaitu *corner*, *dropball*, *goalkick*, kalibrasi, *kickoff*, *stop*, dan tendang. Jadi setiap data suara rekaman akan diklasifikasikan kedalam 2 kelas, yaitu untuk perintah dan orang. Untuk klasifikasi orang, apabila nilai hasil prediksi CNN kurang dari 0.95 maka suara rekaman yang dimasukan akan dianggap berasal dari orang yang tidak dikenal. Sedangkan jika lebih dari 0.95, maka akan diklasifikasikan kedalam suara 5 orang yang dilatihkan.

IV. HASIL PENGUJIAN

A. *Pengujian Rekaman Suara dengan VAD*

Pengujian ini dilakukan untuk merekam suara perintah pada sinyal suara yang ditangkap oleh mikrofon. Hasil rekaman suara dengan VAD dapat dilihat pada Gambar 10.



(a) (b)

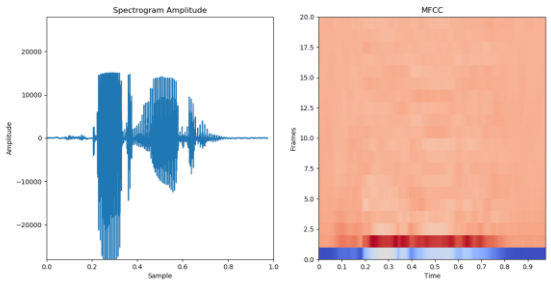
Gambar 10. (a)Sinyal Suara dari Mikrofon (b) Hasil Rekaman dengan VAD

Jadi *voice activity detector* (VAD) akan memulai merekam sinyal suara yang diterima mikrofon ketika sinyal yang diterima melebihi nilai ambang batas yang ditentukan

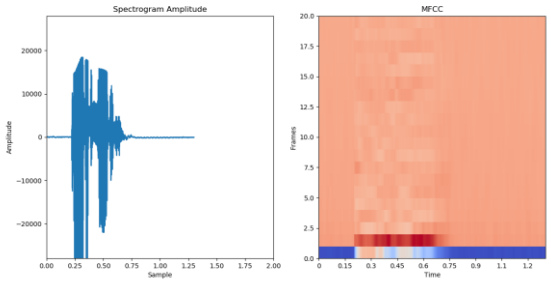
B. *Hasil Fitur MFCC*

Hasil yang didapatkan dari proses perhitungan MFCC pada data rekaman suara yang digunakan untuk proses pelatihan CNN dapat dilihat pada Gambar 11 sampai Gambar 15. Pada gambar-gambar tersebut ditampilkan sinyal hasil rekaman dan gambar hasil ekstraksi fitur MFCC yang didapat dari suara 5 orang laki-laki yang dilatihkan.

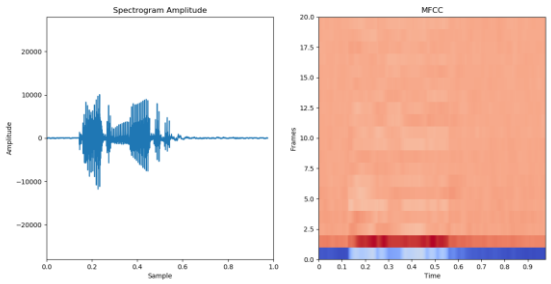
Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*



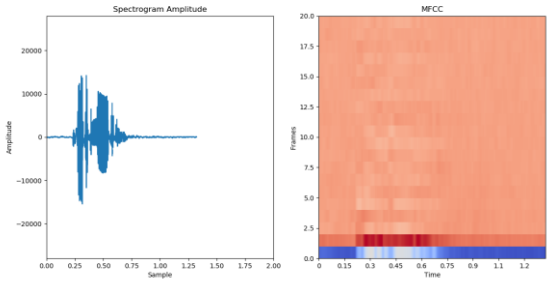
Gambar 11. Fitur MFCC Perintah *Corner* (Arif)



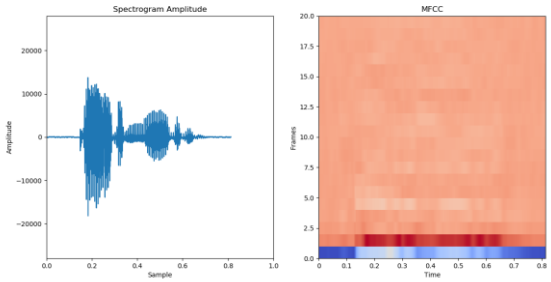
Gambar 12. Fitur MFCC Perintah *Corner* (Azhar)



Gambar 13. Fitur MFCC Perintah *Corner* (Habib)



Gambar 14. Fitur MFCC Perintah *Corner* (Nisar)



Gambar 15. Fitur MFCC Perintah *Corner* (Revo)

C. Pengujian Pengenalan Suara

Pengujian ini dilakukan pada suara 9 orang berbeda, 3 orang merupakan orang yang suaranya dilatihkan pada CNN, sedangkan 6 orang lainya tidak dilatihkan. Suara dari 6 orang yang diujikan terdiri dari 3 orang laki-laki dan 3 orang perempuan. Hasil pengujian untuk suara orang yang dilatihkan dapat dilihat pada Tabel 1. Sedangkan untuk orang yang tidak dilatihkan dapat dilihat pada Tabel 2.

Tabel 1. Hasil Pengujian pada Suara yang Dilatihkan

Nama	Jenis Kelamin (L/P)	Jarak (cm)	Keberhasilan (%)	
			Perintah	Orang
Arif	L	5	100	71.4
		25	100	51.4
		50	97.1	68.6
		100	88.6	42.9
Azhar	L	5	97.1	77.1
		25	100	88.6
		50	97.1	82.9
		100	100	2.9
Nisar	L	5	100	48.6
		25	80	91.4
		50	91.4	82.9
		100	80	80

Tabel 2. Hasil Pengujian pada Suara yang Tidak Dilatihkan

Nama	Jenis Kelamin (L/P)	Jarak (cm)	Keberhasilan (%)	
			Perintah	Orang
Jauhar	L	5	74.3	71.4
		25	100	62.9
		50	97.143	62.9
		100	94.3	85.7
Muklis	L	5	100	88.5
		25	74.3	57.1
		50	88.6	45.7
		100	57.1	60
Mukramin	L	5	100	68.5
		25	94.3	34.3
		50	82.9	54.3
		100	71.4	62.9
Azza	P	5	68.6	68.6
		25	57.1	77.1
		50	51.4	82.9
		100	45.7	48.6
Melvy	P	5	40	80
		25	42.9	91.4
		50	51.4	80
		100	54.3	65.7
Yani	P	5	82.9	71.4
		25	60	77.1
		50	65.7	68.6
		100	57.143	68.6

Hasil akurasi keberhasilan didapatkan pada tabel diatas merupakan nilai rata-rata keberhasilan dari 7 perintah yang diucapkan. Sedangkan untuk pengujian penegenalan orang pada orang yang tidak dilatihkan, dikatakan berhasil apabila suara orang-orang tersebut diklasifikasikan kedalam kelas orang tidak dikenal.

KESIMPULAN

Sistem pengenalan ini dapat diterapkan untuk mengenali suara laki-laki. Berdasarkan data pengujian yang didapatkan, dapat dilihat bahwa sistem pengenalan suara perintah ini memiliki kemampuan mengenali suara laki-laki lebih baik daripada mengenali suara perempuan. Jadi dari hasil pengujian diatas, rata-rata akurasi keberhasilan terbaik pada suara yang dilatihkan adalah sebesar 95% untuk pengenalan perintah dan 78% untuk pengenalan orang. Untuk suara yang tidak dilatihkan didapatkan akurasi keberhasilan sebesar 91% untuk pengenalan perintah dan 76% untuk pengenalan orang pada laki-laki, sedangkan perempuan 64% untuk pengenalan perintah dan 73% untuk pengenalan orang. Penyebab sistem pengenalan suara ini kurang efektif untuk suara perempuan adalah karena suara yang dilatihkan hanya digunakan suara laki-laki saja.

DAFTAR PUSTAKA

[1] W. Q. Ong and A. W. C. Tan, "Robust voice activity detection using gammatone filtering and entropy," in *2016 International Conference on Robotics, Automation and Sciences (ICORAS)*, Melaka, Malaysia, 2016, pp. 1–5.

[2] G. Meoni, L. Pilato, and L. Fanucci, "A low power Voice Activity Detector for portable applications," in *2018 14th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, Prague, 2018, pp. 41–44.

[3] T. Nasution, "Metoda Mel Frequency Cepstrum Coefficients (MFCC) untuk Mengenali Ucapan pada Bahasa Indonesia," vol. 1, no. 1, p. 10, 2012.

[4] M. Bezoui, A. Elmoutaouakkil, and A. Beni-hssane, "Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coeficients (MFCC)," in *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*, Marrakech, Morocco, 2016, pp. 127–131.

[5] A. Winursito, R. Hidayat, A. Bejo, and M. N. Y. Utomo, "Feature Data Reduction of MFCC Using PCA and SVD in Speech Recognition System," in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, Shah Alam, 2018, pp. 1–6.

[6] W. S. Eka Putra, "Klasifikasi Citra Menggunakan Convolutional Neural Network (CNN) pada Caltech 101," *J. Tek. ITS*, vol. 5, no. 1, Mar. 2016.

[7] E. N. Arrofiqoh and H. Harintaka, "Implementasi Metode Convolutional Neural Network Untuk Klasifikasi Tanaman Pada Citra Resolusi Tinggi," *GEOMATIKA*, vol. 24, no. 2, p. 61, Nov. 2018.