RANCANG BANGUN KUNCI BERBASIS SUARA PADA PINTU PINTAR DENGAN MENGGUNAKAN METODE MEL FREQUENCY CEPSTRAL COEFFICIENT (MFCC) DAN K-NEAREST NEIGHBOR (K-NN)

DESIGN OF SMART DOOR KEY SYSTEM BASED ON VOICE RECOGNITION USING MEL FREQUENCY CEPSTRAL COEFFICIENT (MFCC) AND K-NEAREST NEIGHBOR (K-NN)

Muhammad Afif Ridwansyah[1], Achmad Rizal, S.T., M.T.[2], Sugondo Hadiyoso, ST., MT.[3]

[1,2,3]Prodi S1 Teknik Elektro, Fakultas Teknik Elektro, Universitas Telkom

[1] m.afif.rid@gmail.com, [2] achmadrizal@telkomuniversity.ac.id, [3] sugondo@ telkomuniversity.ac.id

Abstrak

***Automatic Speech Recognition* (ASR) adalah suatu sistem yang dapat mengenali, membandingkan dan mencocokkan pola suara masukan sistem tersebut dengan pola suara yang telah disimpan dalam memori. ASR terbagi menjadi dua jenis, yaitu *Speech Recognition* dan *Speaker Recognition*. *Speaker Recognition* adalah pengenalan identitas berdasarkan suara yang dikeluarkan (berupa intonasi suara, kedalaman suara, dan sebagainya). Pada penelitian ini dibangun sistem kunci berbasis suara dengan memanfaatkan *Speaker Recognition*. Pada penelitian ini digunakan metode *Mel-Frequency Cepstral Coefficient* (MFCC) sebagai ekstraksi ciri dan metode *K-Nearest Neighbor* (K-NN) sebagai klasifikasi ciri. Alat ini bekerja melalui dua tahapan, yaitu tahap pelatihan (*training*) dan tahap pengujian (*testing*). Hasil pengujian menunjukkan MFCC dan K-NN berhasil diimplementasikan dengan jumlah *filterbank* terbaik berjumlah 20 dan nilai koefisien terbaik sebanyak 13 koefisien dengan akurasi 100%. Hasil pengujian menunjukkan bahwa jumlah *filterbank* dan nilai koefisien mempengaruhi akurasi dari sistem.**

Kata kunci: *Automatic Speech Recognition* (ASR), biometrik suara, *K-Nearest Neighbor* (K-NN), *Mel-Frequency Cepstral Coefficient* (MFCC) .

Abstract

Automatic Speech Recognition (ASR) is a system that can identify, compare and match the system input voice patterns with the voice patterns that has been stored in memory. ASR is divided into two types, namely Speech Recognition and Speaker Recognition. Speaker Recognition is the introduction of the issued voice character (intonation of sound, depth of voice, etc.). The key system based on voice using Speaker Recognition was build in this study. In this research, the methods used were Mel-Frequency Cepstral Coefficient (MFCC) as feature extraction and K-Nearest Neighbor (K-NN) as characteristic classification. This tool worked through two stages, namely training stage and testing stage. The results showed that the MFCC and K-NN were successfully implemented with best filter bank at number 20 filter, best coefficient value at 13 coefficient with 100% accuracy. The results showed that filter bank and coefficient affect the accuracy of the system.

Keywords : Automatic Speech Recognition (ASR) , Biometric, K-Nearest Neighbor (K-NN), Mel-Frequency Cepstral Coefficient (MFCC).

1. PENDAHULUAN

Teknologi pada sistem keamanan sudah semakin berkembang pesat, mulai dari penggunaan kunci, pin, kata sandi dan kode batang (*barcode*). Namun metode ini memiliki kekurangan seperti mudah dilupakan, diretas(*hack*) atau diduplikat. Salah satu teknologi yang sekarang sedang banyak dikembangkan adalah teknologi biometrik. Teknologi ini adalah suatu sistem pengenalan individu berdasarkan ciri khas atau sifat biologis yang dimiliki oleh seseorang, seperti pola retina (berdasarkan pola pada pembuluh darah pada mata), karakteristik muka (berdasarkan letak anggota tubuh pada muka), sidik jari, dan bentuk gigi[1]. Teknologi ini lebih baik dibandingkan teknologi sebelumnya karena identifikasi menggunakan teknologi biometrik menyingkirkan kebutuhan mengingat kata sandi serta pin atau membawa token seperti kunci atau kartu.

Seiring dengan perkembangan ilmu pengetahuan, kemudian dilakukan pengidentifikasian individu melalui suara, sebagai contoh pengenalan suara sebagai fungsi membuka kunci pada *handset* android[2]. Suara manusia merupakan kombinasi kompleks dari variasi tekanan udara dari paru-paru yang melewati pita suara dan saluran vokal. Hal ini menghasilkan karakteristik unik pada suara sehingga suara setiap individu berbeda dan sulit untuk ditiru. Faktor yang membuat setiap individu menghasilkan suara yang berbeda-beda adalah saluran vokal, seperti: hidung, lidah, bibir dan gigi[3].

Automatic Speech Recognition (ASR) adalah suatu sistem yang dapat mengenali, membandingkan dan mencocokkan pola suara masukan sistem tersebut dengan pola suara yang telah disimpan dalam memori[4]. *Automatic Speech Recognition* terbagi menjadi dua jenis, yaitu Pengenalan Kata (*Speech Recognition*) dan Pengenalan Suara (*Speaker Recognition*). *Speaker Recognition* adalah pengenalan identitas berdasarkan suara yang dikeluarkan (berupa intonasi suara, kedalaman suara, dan sebagainya). *Speech Recognition* adalah proses dimana seseorang memberikan masukan suara kepada komputer, kemudian komputer mengenali kata yang diucapkan seseorang.

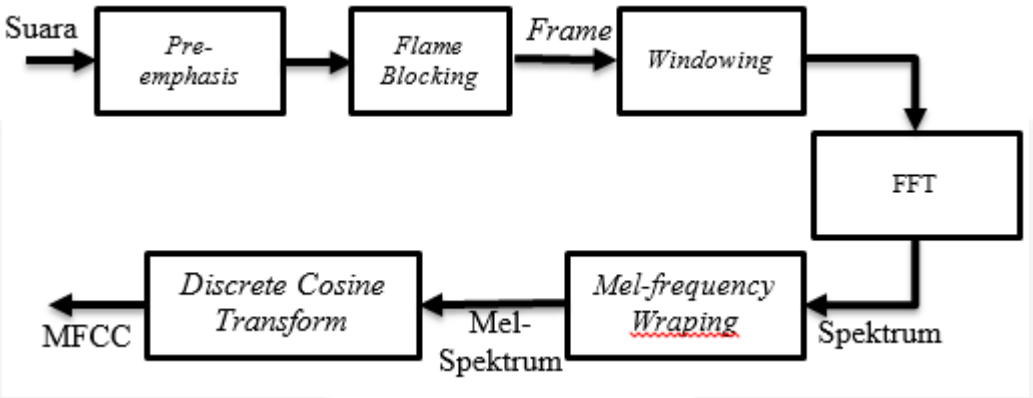
Dengan memanfaatkan *Speaker Recognition*, akan dibangun sistem kunci berbasis suara dengan mengkombinasikan metode *Mel Frequency Cepstral Coefficient* (MFCC) sebagai ekstraksi cirinya dan *K-Nearest Neighbor* (K-NN) untuk klasifikasinya. Alat ini bekerja melalui dua tahapan, yaitu tahap pelatihan (training) dan tahap pengujian (*testing*). Tahapan pertama, tahapan pengenalan sistem, dalam tahapan ini sistem diberi kemampuan untuk mengenali karakteristik yang akan menjadi *database* (tahap latih). Tahap kedua sistem akan diberikan masukkan berupa suara, kemudian suara tersebut dicocokan dengan *database* (tahap uji)[5].

2. TINJAUN PUSTAKA DAN PERANCANGAN

2.1 *Mel Frequency Cepstrum Coefficients* (MFCC)

Mel- Frequency Cepstrum Coefficients (MFCC) merupakan metode ekstaksi ciri yang digunakan dalam pengolahan suara. Metode ini digunakan untuk mengkonversikan sinyal suara menjadi beberapa parameter.

MFCC didasarkan atas persepsi ambang dengar dan batas sakit telinga manusia, MFCC memiliki dua jenis filter yang merupakan spasi linier pada frekuensi rendah di bawah 1000 Hz dan logaritmik di atas 1000 Hz[4]. Diagram blok dari pemroses MFCC dapat dilihat pada Gambar 1.



Gambar 1. Diagram blok pemroses MFCC.

1. Pre-emphasis

Pre-emphasis adalah langkah yang memproses sinyal melalui filter yang akan meningkatkan energi sinyal pada frekuensi yang lebih tinggi dan meratakan daya sinyal. Hal ini disebabkan sinyal suara dengan frekuensi tinggi memiliki magnitude atau energi yang lebih kecil dibandingkan dengan sinyal berfrekuensi rendah[4]. Persamaan yang digunakan dapat dilihat pada persamaan (1)

$$Y[n] = X[n] - a\,X[n-1] \qquad a = 0{,}97 \qquad \ldots(1)$$

2. Frame Blocking

Proses *frame blocking* merupakan proses memecah sinyal menjadi segmen-segmen yang disebut dengan frame, sehingga didapat karakter sinyal yang lebih stabil. Hal ini dilakukan karena sinyal suara manusia merupakan sinyal yang tidak stabil, dimana intensitas dan kuat amplitudo suara selalu berubah[10]. Pada langkah ini, sinyal suara yang terdiri dari S sampel (X(S)) dibagi menjadi beberapa *frame* yang berisi N sampel, masing-masing *frame* dipisahkan oleh M (M<N) dengan panjang frame dalam kisaran 20ms sampai 40ms. *Frame* pertama berisi sampel N pertama. *Frame* kedua dimulai M sampel setelah permulaaan *frame* pertama, sehingga *frame* kedua ini *overlap* terhadap *frame* pertama sebanyak N-M sampel. Seterusnya, *frame* ketiga dimulai M sampel setelah *frame* kedua (juga *overlap* sebanyak N-M sampel terhadap *frame* kedua). Proses ini berlanjut sampai seluruh suara tercakup dalam *frame*. Hasil dari proses ini adalah matriks dengan N baris dan beberapa kolom sinyal X[N].

3. Windowing

Langkah selanjutnya adalah *windowing* setiap *frame* untuk meminimalisir diskontinuitas sinyal pada permulaan dan akhir setiap *frame*. Konsepnya adalah meruncingkan sinyal ke angka nol pada permulaan dan akhir setiap *frame*[4]. Bila *window* didefinisikan sebagai w(n), $0 \leq n \leq N-1$, dengan N adalah jumlah sampel dalam tiap *frame*, maka hasil dari proses ini adalah sinyal :

$$y(n) = x(n)w(n), 0 \leq n \leq N-1 \quad \ldots(2)$$

Keterangan:

y(n) = sinyal hasil *windowing* sampel ke-n
x(n) = nilai sampel ke-n
w(n) = nilai *window* ke-n
N = jumlah sampel dalam *frame*

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right) \quad \ldots(3)$$

4. Fast Fourier Transform (FFT)

Proses selanjutnya adalah mengkonversi setiap *frame* dari ranah waktu ke ranah frekuensi. *Discrete Fourier Transform* (DFT) merupakan prosedur matematika yang digunakan untuk menentukan frekuensi yang merupakan isi dari urutan sinyal diskrit. DFT berasal dari fungsi *Fourier Transform* (FT) yang didefinisikan pada persamaan (4)

$$X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft}\, dt \quad \ldots(4)$$

Keterangan :

X(f) = Sinyal dalam ranah frekuensi
x(t) = Sinyal dalam ranah waktu

Dengan munculnya komputer digital, ilmuan di bidang pengolahan digital berhasil mendefinisikan DFT sebagai berikut :

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}} \quad k = 0,1,2,\ldots,N-1 \quad \ldots(5)$$

Keterangan :

X(k) = Output DFT
N = Jumlah sampel yang akan diproses
x(n) = Nilai sampel sinyal
k = Variabel frequensi diskrit

Namun, DFT sangat tidak efisien karena jumlah titik dalam DFT berjumlah ribuan, sehinggal jumlah yang dihitung tidak dapat ditentukan. Pada tahun 1965, COOLEY dan TUKEY menjelaskan algoritma yang sangat efisien untuk menerapkan DFT, yang disebut dengan *Fast Fourier Transform* (FFT). FFT dipergunakan untuk mengurangi kompleksitas transformasi yang dilakukan oleh DFT. Sebagai perbandingan, misalkan jumlah sampel (N) yang diambil sebanyak 2 sampel, tingkat kompleksitas DFT adalah $N^2 = 4$, sedangkan kompleksitas FFT adalah N log N = 0,602.

5. Mel-Frequency Wrapping

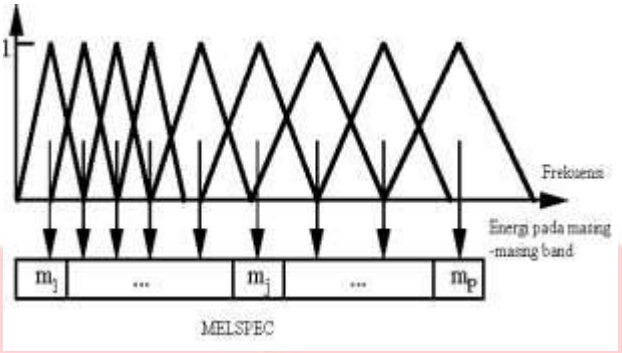
Studi psikofisik telah menunjukkan bahwa persepsi manusia tentang frekuensi suara untuk sinyal ucapan tidak mengikuti skala linier. Jadi, untuk setiap nada dengan frekuensi sesungguhnya f (satuan dalam Hz), sebuah pola diukur dalam sebuah skala yang disebut "mel". Skala "mel frekuensi" adalah skala frekuensi linier di bawah 1000 Hz dan skala logaritmik di atas 1000 Hz[10]. Skala ini didefinisikan oleh Stanley Smith, John Volkman dan Edwin Newman sebagai :

$$F(Mel) = \left[2595 * \log 10\left(1 + \frac{f}{700}\right)\right] \quad \ldots(6)$$

Keterangan :

F(Mel) = Fungsi Mel Scale
f = Frekuensi

Sebuah pendekatan untuk simulasi spektrum dalam skala mel adalah dengan menggunakan *filter bank* yang diletakkan secara seragam dalam skala mel yang ditunjukkan pada Gambar 2.



Gambar 2. Contoh *mel-spaced filter bank*.

6. **Discrete Cosine Transform (DCT)**

Pada langkah terakhir ini, spektrum log mel dikonversi menjadi ranah waktu menggunakan *Discrete Cosine Transform* (DCT). Hasil dari proses ini dinamakan *Mel-Frequency Cepstrum Coefficients* (MFCC). Kumpulan dari koefisien ini disebut sebagai vektor akustik yang akan digunakan sebagai nilai yang mewakili sinyal suara.

$$C_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad \ldots (7)$$

Keterangan :

S_k = *mel power spectrum coefficients*

K = Jumlah koefisien yang diharapkan

2.2 *K-Nearest Neighbor* (K-NN)

Dalam sistem pengenalan pola, *k-nearest neighbor* adalah metode yang nonparametik yang digunakan untuk klasifikasi dan regresi. Klasifikasi tidak menggunakan model apapun, hanya dengan menemukan titik terdekat dari data pelatihan.

1. ***Euclidean Distance***

Euclidean Distance $d_E(x,y)$ merupakan persamaan yang digunakan untuk menghitung jarak antara pelatihan dan pengujian sinyal suara terdiri dari fitur N, yaitu $x = \{x_1,x_2,\ldots,x_N\}$, $y = \{y_1,y_2,\ldots,y_N\}$. *Euclidean distance* didefinisikan dalam persamaan (8).

$$d_E(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad \ldots (8)$$

Keterangan :

$d_E(x,y)$ = Jarak skalar antara dua buah vektor x dan y dari matriks D dimensi

i = Jumlah data ke n

N = Jumlah data

x = Data pelatihan

y = Data pengujian

2.3 Perancangan dan Realisasi Sistem

Berikut adalah perangkat yang dibutuhkan dalam pembuatan alat:

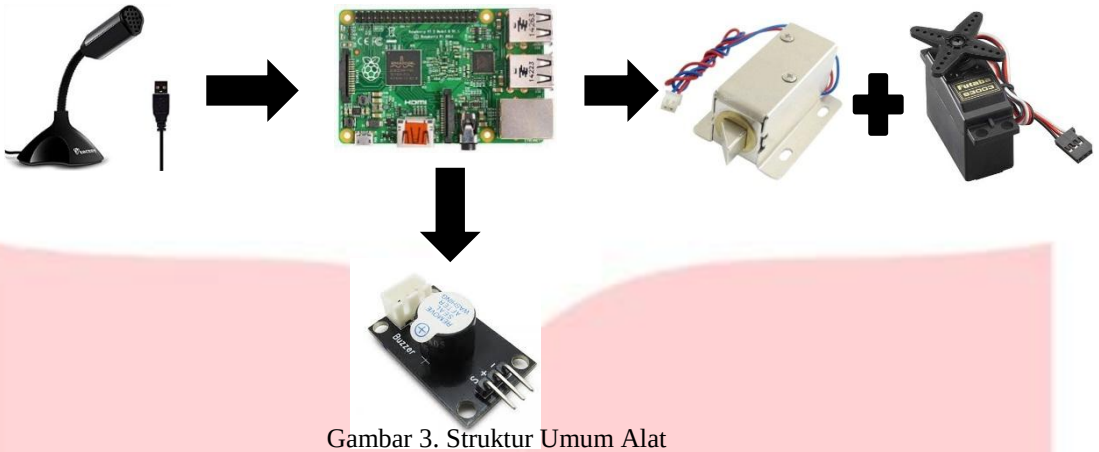
1) *Raspberry pi 2 model B*

2) Mikrofon dengan *port* USB

3) *Buzzer*

4) Kunci Solenoid

5) Motor Servo



Gambar 3. Struktur Umum Alat

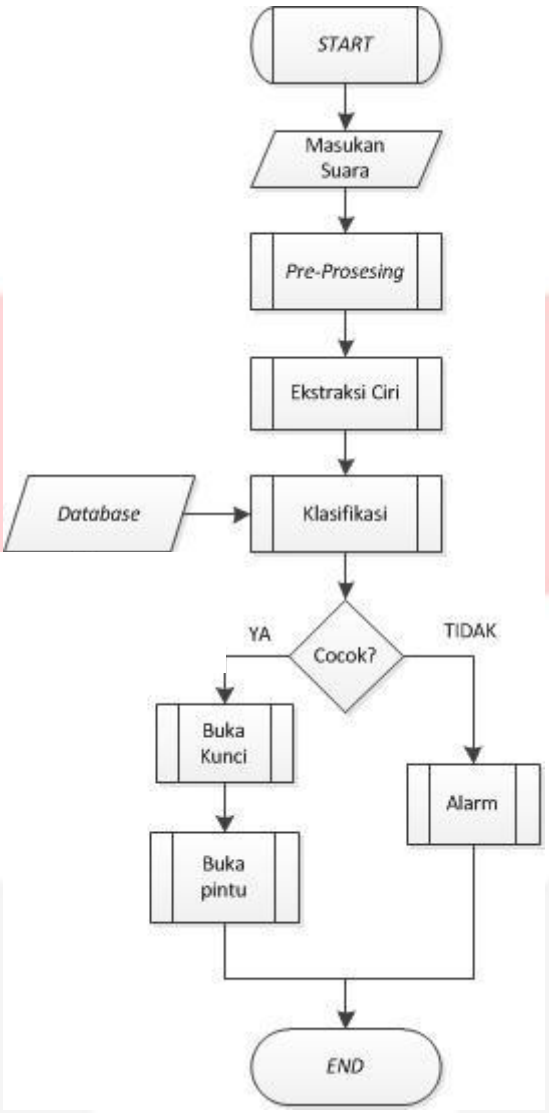
Pada Gambar 3, suara akan diambil menggunakan mikrofon dengan *port* USB yang akan menjadi masukan untuk *raspberry pi 2 model B*. Pada tahap selanjutnya akan dilakukan proses pengolahan suara pada *raspberry pi 2 model B* yang hasilnya akan menjadi masukan untuk kunci solenoid, motor servo dan *buzzer*.

Cara kerja dari alat ini terbagi menjadi 2 fase yaitu fase pelatihan dan fase pengujian. Fase pelatihan merupakan tahan pembelajaran dan pemodelan suara yang nantinya akan simpan ke dalam *database*. Fase pengujian merupakan tahap untuk mengenali suara yang diterima mikrofon apakah sesuai dengan *database* atau tidak. Jika sesuai dengan *database*, kunci dan pintu akan terbuka. Jika tidak sesuai, alarm akan berbunyi.

2.4 Diagram Alir Sistem

Gambar 4. Diagram Alir Fase Pelatihan

Gambar 4 menampilkan diagram alir fase pelatihan. Pada fase pelatihan akan dilakukan proses *pre-processing* pada sinyal suara berupa normalisasi dan *silence removal*. Selanjutnya hasil dari *pre-processing* akan di ekstraksi ciri menggunakan metode MFCC untuk mengambil informasi penting pada sinyal. Hasil dari ekstraksi ciri akan disimpan dalam file *database*. Data pada file *database* ini yang nantinya akan dijadikan perbandingan dengan sinyal suara pada fase pengujian.



Gambar 5. Diagram Alir Fase Pengujian

Gambar 5 menampilkan diagram alir fase pengujian. Pada fase ini akan dilakukan proses *pre-processing* pada sinyal suara berupa normalisasi dan *silence removal*. Selanjutnya akan di ekstraksi ciri melalui metode MFCC untuk mengambil informasi penting pada sinyal. Hasil dari ekstraksi ciri pada fase pengujian dibandingkan dengan seluruh data hasil ekstraksi ciri yang disimpan pada *database* pada fase pelatihan. Proses klasifikasi tersebut menggunakan metode K-NN, dimana jika hasilnya sesuai maka kunci dan pintu akan terbuka. Jika tidak maka alarm akan berbunyi.

3. PENGUJIAN DAN ANALISIS

3.1 Pengujian dan Analisis Parameter

Pada penelitian ini dilakukan analisi akurasi berdasarkan jumlah *filterbank* dan jumlah koefisien yang diterapkan pada sistem. Menurut teori jumlah *filterbank* yang biasa digunakan adalah pada kisaran 20-40. Pada pengujian parameter ini diambil jumlah filterbank sebanyak 20, 30 dan 40 sebagai nilai pengujian. Untuk jumlah koefisien yang diambil adalah 8, 13, 26.

Tabel 1. Akurasi berdasarkan skenario ke-1 (*filterbank* = 20 filter)

Koefisien	Jumlah Percobaan	Dikenali	Tidak Dikenali	Akurasi (%)
8	25	22	3	88%
13	25	25	0	100%
26	25	23	2	92%

Tabel 2. Akurasi berdasarkan skenario ke-2 (*filterbank* = 30 filter)

Koefisien	Jumlah Percobaan	Dikenali	Tidak Dikenali	Akurasi (%)
8	25	17	8	68%
13	25	21	3	84%
26	25	19	6	76%

Tabel 3.Akurasi berdasarkan skenario ke-3 (*filterbank* = 40 filter)

Koefisien	Jumlah Percobaan	Dikenali	Tidak Dikenali	Akurasi (%)
8	25	15	10	60%
13	25	20	5	80%
26	25	17	8	68%

Berdasarkan Tabel 1, dapat dilihat bahwa jumlah filterbank terbaik adalah 20 dan jumlah koefisien terbaik adalah 13. Jumlah filterbank berpengaruh terhadap penyaringan hasil FFT. Dapat dilihat semakin banyak jumlah filterbank maka akurasi semakin menurun. Jumlah koefisien berpengaruh terhadap data yang diuji. Terlalu sedikit koefisien tidak cukup untuk mewakili data yang akan dikenali, karena proses pengenalan akan semakin sulit untuk berhasil. Terlalu banyak akan membuat ciri semakin tidak jelas, hal ini disebabkan jika terlalu banyak maka akan membuat sistem bingung karena adanya beberapa titik ciri yang mirip.

3.2 Pengujian dan Analisis *Real Time*

1. Pengujian Identifikasi Identitas

Pengujian dilakukan oleh lima orang yang mempunyai *database*. Pengujian dilakukan untuk melihat apakah si pembicara dapat diidentifikasi oleh sistem.

Tabel 4. Pengujian Identifikasi Identitas

Pembicara	Dikenali sebagai					Jumlah Percobaan
	S1	S2	S3	S4	S5	
S1	40	0	0	0	0	40
S2	0	40	0	0	0	40
S3	0	0	40	0	0	40
S4	0	0	0	40	0	40
S5	0	0	0	0	40	40

Pada Tabel 4. dapat dilihat dari percobaan yang dilakukan bahwa sistem mampu mengidentifikasi si pembicara.

2. Pengujian Sistem

Pengujian dilakukan dengan menunjuk salah satu dari lima orang yang mempunyai *database* sebagai *user*, dimana hanya dia yang mampu mengakses pintu. Dalam penelitian ini ditunjuklah S3 sebagai *user* secara acak dan keempat orang lainnya sebagai *non-user*. Proses pengujian dilakukan dengan dengan 2 cara, pertama proses identifikasi S3 sebagai *user* dan kedua proses identifikasi orang lain yang tidak ada di dalam *database*. Pengujian pertama dilakukan untuk melihat apakah si pembicara di identifikasi sebagai dirinya atau sebagai orang lain. Pengujian kedua dilakukan untuk melihat pengaruh data tidak dikenal pada sistem.

Tabel 5. Pengujian Sistem

Pembicara	Dikenali sebagai					Jumlah Percobaan
	S1	S2	S3	S4	S5	
U	0	0	40	0	0	40
NS	0	0	0	0	40	40

Pada Tabel 5. dapat dilihat bawah sistem dapat mengidentifikasi si pembicara S3 sebagai dirinya sendiri. Sedangkan pembicara asing diidentifikasi sebagai S5, hal ini terjadi karena pada proses klasifikasi K-NN sistem akan membandingkan pembicara asing dengan seluruh data yang ada pada *database* dan memberi *output* data dengan nilai terdekat dengan pembicara asing.

4. KESIMPULAN

Berdasarkan hasil pengujian dan analisis yang telah dilakukan pada sistem ,maka dapat diambil kesimpulan sebagai berikut :

1. Pada pengujian, jumlah filterbank terbaik berjumlah 20 filter dan nilai koefisien terbaik sebanyak 13 koefisien dengan akurasi 100%
2. Banyaknya *filter*, jumlah koefisien pada MFCC dapat mempengaruhi akurasi sistem dalam mengenali perintah.
3. Semakin banyak data yang dibandingkan maka akurasi sistem akan semakin menurun dan proses komputasi akan semakin lama.
4. Dari hasil keseleruhan, sistem mampu menggabungkan metode MFCC dan K-NN dengan kecepatan respon selama 4,3832 detik

DAFTAR PUSTAKA

[1] Pankanti, S. (2015, November). Biometric Recognition : Security and Privacy Concerns. *IEEE Security and Privacy Magazine*, pp. 36.

[2] Natalia, D. (2013). Perancangan dan Implementasi Speech Recognition System Sebagai Fungsi Unlock pada Handset Android. *Pengolahan Sinyal Informasi*.

[3] J. Holmes, &. W. Holmes (2001). Mechanisms and Models of Human Speech Production. In F. Taylon, *Speech Synthesis and Recognition, 2nd ed* (pp. 11-31). USA and Canada.

[4] L. Muda, M. B. (2010). Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. In *Journal of Computing, Vol. 2, No. 3* (pp. 138-143).

[5] N. Baranwal, S. T. (2014). A Speaker Invariant Speech Recognition Technique Using MFCC Features in Isolated Hindi Word. *International Journal of Computational Intelligence Studies, Vol. 3*, 277-291.

[6] Hermanto, H. (2016). *Modul Pelatihan Guru Paket Keahlian Teknik Audio Video.* Malang: PPPPTK.

[7] Upton, E. (2015, February 2). *Raspberry Pi 2 on Sale Now.* Retrieved September 26, 2017, from Raspberry Pi: www.raspberrypi.org/blog/raspberry-pi-2-on-sale

[8] Martin. (2016, February 2). *Fungsi Microphone*. Retrieved September 26, 2017, from Fungsiklopedia.com: http://www.fungsiklopedia.com/fungsi-microphone/

[9] Yagusandri, A. (2011). *Rancang Bangun Prototipe Sistem Aktuator Sirip Roket Menggunakan Motor Servo*. Depok: Universitas Indonesia.

[10] *Filterbanks, Mel-Frequency Cepstral Coefficients (MFCC) and What's In-Between*. (2016, 04 21). Retrieved September 26, 2017, from Speech Processing for Machine Learning: www.haythamfayek.com

[11] Anil Jail, L. H. (2000). Biometric Identification. In *Communications of The ACM, Vol. 43, No. 2* (p. 93).

[12] A. Bombatkar, G. B. (2014). Emotion Recognition using Speech Recognition using K-Nearest Neighbor Algorithm. *International Journal of Engineering Research and Applications (IJERA)*, pp. 2248-9622.