APLIKASI PENGENAL PENGUCAP BERBASIS IDENTIFIKASI SUARA DENGAN EKSTRAKSI CIRI MEL-FREQUENCY CEPSTRUM COEFFICIENTS (MFCC) DAN KUANTISASI VEKTOR

Mega Tiara Nur Azizah[*)], Achmad Hidayatno, and Yuli Christyono

Departemen Teknik Elektro, Universitas Diponegoro
Jl. Prof. Sudharto, SH, Kampus UNDIP Tembalang, Semarang 50275, Indonesia

[*)]*E-mail: tiaramega26@gmail.com*

Abstrak

Kemajuan teknologi dalam bidang Pengolahan Sinyal Digital telah berkembang pesat dan membawa dampak positif dalam kehidupan manusia. Salah satu disiplin ilmu dalam pengolahan sinyal digital yang memberikan dampak yang cukup besar ialah bidang Pengolahan Suara Digital dan salah satu pengembanganya adalah pengenalan pengucap. Pengenalan pengucap dapat digunakan untuk sistem keamanan, absensi dan lain sebagainya. Program pengenal pengucap ini menggunakan ekstraksi ciri Mel Frequency Cepstrum Coefficient (MFCC) dan Kuantisasi Vektor untuk menghasilkan koefisien-koefisien ciri dari masing-masing suara responden. Dengan menghitung jarak Euclidean dan jarak Mahalanobis terdekat maka akan diambil keputusan atas kepemilikan suara pengucap. Apabila hasil keputusan dengan menggunakan jarak Euclidean dan jarak Mahalanobis sama atau match maka suara pengucap tersebut akan dikenali sedangkan apabila hasil keputusan antara kedua jarak tersebut berbeda maka pengucap tidak akan dikenali. Pengujian dilakukan dalam 3 variasi yaitu variasi sample rate, ukuran codebook, dan kondisi tidak ideal/salah. Hasil pengujian pada variasi sample rate didapat akurasi tertinggi pada saat sample rate bernilai 16000Hz yaitu sebesar 83,3%, sedangkan pada variasi ukuran codebook didapat akurasi tertinggi pada saat ukuran codebook 16 dan hasil pengujian dengan kondisi tidak ideal/salah didapatkan akurasi 100%.

Kata kunci : MFCC, kuantisasi vektor, jarak Mahalanobis, jarak Euclidean, pengenalan pengucap.

Abstract

Technology advances in Digital Signal Processing sector is rapidly developing and brings positive impacts into human's life. One of study disciplines in digital signal processing that brings significant impact is the Digital Voice Processing sector and one of the development is speaker recognition. Speaker recognition can be used for security system, attendance and many more. This Speaker Recognition program is using Mel Frequency Cepstrum Coefficient (MFCC) characteristic extraction and Vector Quantity to generate characteristic coefficients from each respondent's speech. By calculating the closest Euclidean range and Mahalanobis range, decision of speech's voice ownership will be taken. If the decision using Euclidean range and Mahalanobis range is same or matched then the speech will be recognized, while if the decision of those two range is different then the speech will not be recognized. The program is tested in 3 variations that is sample rate, codebook size and not ideal/wrong conditions variations. Testing result in sample rate variation is obtained that the higher accuracy is when the sample rate is 16000 Hz as much as 83,3%. While in the codebook size variation highest accuracy is obtained when the codebook size is 16 and the not ideal/wrong condition variation testing has 100% accuracy

Keywords: MFCC, vector quantity, Mahalanobis range, Euclidean range, speaker recognition.

1. Pendahuluan

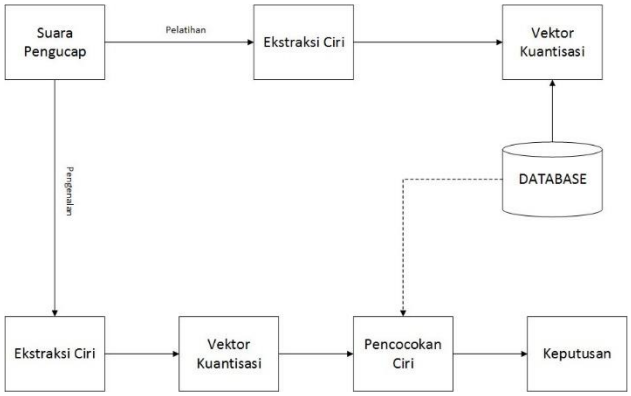
Ilmu pengetahuan dan teknologi khususnya pengolahan sinyal memegang peranan yang penting. Penelitian yang intensif dalam bidang pengolahan sinyal menyebabkan teknologi komunikasi berkembang dengan pesat. Salah satunya adalah pengenalan pengucap. Pengenalan pengucap merupakan cara yang digunakan untuk mengetahui identitas seseorang yang mengucapkan sinyal informasi. Ucapan berisi beberapa karakteristik yang spesifik untuk setiap individu, yang beberapa diantaranya tidak dipengaruhi oleh pesan linguistik yang terkandung dalam suatu ucapan[6]. Perbedaan karakteristik ucapan itulah yang menjadi dasar pengenalan pengucap melalui ucapannya.

Proses pengenalan pengucap dalam mengetahui siapa yang mengucapkan sinyal informasi tersebut dengan mencocokkan karakteristik ucapan yang ada di dalam basisdata dengan ucapan masukan. Karakteristik ucapan dapat dibedakan melalui ekstraksi dengan suatu teknik pengkodean. Pada Penelitian ini akan dibahas mengenai analisis dan simulasi sistem pengenalan pengucap menggunakan metode MFCC untuk ekstraksi ciri dari sinyal ucapan yang diinputkan dan jarak Mahalanobis dan jarak Euclidean sebagai penentuan jarak minimum dalam proses pencocokan ciri dari suatu sinyal ucapan. Parameter MFCC dipilih karena parameter ini dapat menyederhanakan kandungan sinyal suara ke dalam koefisien cepstral, lebih tepatnya dipetakan terhadap koefisien mel yang mempunyai respon frekuensi liniear untuk frekuensi kurang dari 1khz, hal ini seperti karakter respon frekuensi pada telinga manusia[5].

2. Metode

2.1. Algoritma dan Diagram Sistem

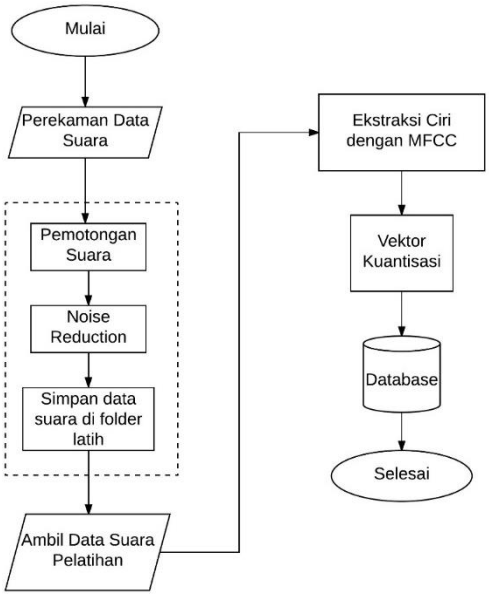
Sistem yang dirancang adalah perangkat lunak guna mengenali suara pengucap yang terdiri dari proses pelatihan dan pengujian.



Gambar 1. Diagram blok pengenalan pengucap.

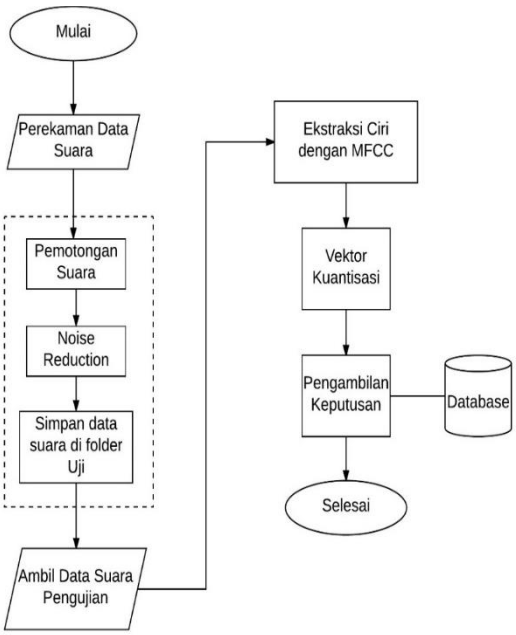
Diagram perancangan dari sistem yang dibuat ditampilkan pada Gambar 1, diagram perancangan ini berfungsi sebagai umpan balik dan sebagai pengawasan terhadap program perangkat lunak selain itu diagram perencanaan dibuat guna mengetahui proses pengolahan dari mengekstraksi ciri suara responden sampai dengan pengambilan keputusan apakah suara responden dikenali atau tidak.

Sistem terdiri dari dua proses yang pertama adalah proses pelatihan yang mana pada proses ini akan dihasilkan database vektor ciri dari setiap data sampel. Proses pelatihan terdiri dari proses perekaman suara yang disimpan kedalam folder data latih, kemudian proses ekstraksi ciri dengan MFCC, dan proses pemetaan vektor dengan algoritma vektor kuantisasi LBG. Diagram alir proses pelatihan ditampilkan oleh Gambar 2.



Gambar 2. Diagram alir proses pelatihan.

Proses yang kedua adalah proses pengujian. Proses pengujian ini memilki tahap yang sama seperti pada tahap pelatihan yaitu proses perekaman suara kemudian ekstraksi ciri dengan MFCC dan vektor kuantisasi untuk mendapatkan codebook, akan tetapi tidak menghasilkan database melainkan menghasilkan keputusan yang diambil dari perbandingan jarak terdekat data yang diuji dengan database yang sudah dimiliki pada proses pelatihan. Diagram alir proses pelatihan ditampilkan oleh Gambar 3.



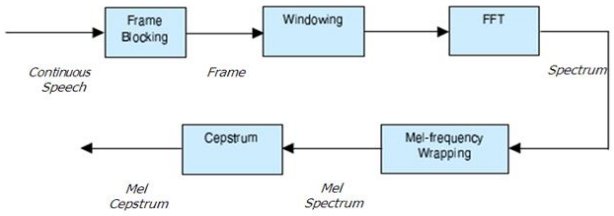
Gambar 3. Diagram alir proses pengujian.

2.2. Ekstraksi Ciri

Ekstraksi ciri atau feature extraction merupakan proses yang mana tiap-tiap sampel sinyal akan diubah menjadi vektor-vektor data. Metode yang akan digunakan pada proses ekstraksi dalam pengenalan pengucap ini adalah dengan menggunakan *Mel Frequency Cepstrum Coefficient* (MFCC). MFCC memiliki beberapa keunggulan dibandingkan dengan metode lainnya, antara lain [9]:

1. Mampu menangkap informasi-informasi penting yang terkandung dalam sinyal suara.
2. Menghasilkan data seminimal mungkin, tanpa menghilangkan informasi-informasi penting yang ada.
3. Mengadaptasi organ pendengaran manusia dalam melakukan persepsi terhadap sinyal suara.
4. Menghasilkan pendekatan yang lebih baik terhadap sistem pendengaran manusia karena menggunakan fungsi logaritmik dalam perhitungannya.

MFCC pada sistem ini bertujuan untuk menghasilkan cepstrum yang akan digunakan dalam membentuk codeword. Blok diagram dari MFCC ditunjukkan pada gambar 4 berikut.



Gambar 4 Proses MFCC.

Proses MFCC diawali dengan membagi sinyal suara menjadi beberapa frame melalu proses frame blocking. Setelah itu dilakukan windowing pada setiap frame. Windowing bertujuan untuk meminimalisasi diskontinuitas sinyal dan distorsi spektral. Masing-masing frame dicari spektrum amplitudonya dengan terlebih dahulu merubah setiap frame dari domain waktu ke domain frekuensi dengan menggunakan Fast Fourier Transform (FFT). Selanjutnya dilakukan proses mel-frequency wrapping untuk memperoleh sinyal spektrum dalam mel-scale dari hasil FFT. Langkah terakhir adalah mengubah hasil log mel spectrum ke dalam domain waktu dan menghasilkan MFCC sebagai hasil akhir. Urutan dan cara kerja MFCC dapat dijelaskan sebagai berikut [11].

1. Frame Blocking

Sinyal suara merupakan sinyal yang tidak stabil sehingga tidak dapat dilakukan ekstraksi ciri secara langsung. Oleh karena itu perlu dilakukan proses frame blocking dengan membagi sinyal suara menjadi sejumlah N-frame. Pada langkah ini, sinyal ucapan yang terdiri dari n sampel (x(n)) dibagi menjadi beberapa frame yang berisi N sample, masing-masing frame dipisahkan oleh M (M<N). Frame pertama berisi sampel N pertama. Frame kedua dimulai M sampel setelah permulaan frame pertama, sehingga frame kedua 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).

2. Windowing

Windowing dilakukan untuk memperkecil penyimpangan pada sinyal yang diskontinu di awal dan di akhir masing-masing frame. Ada banyak jenis window, misalnya hamming, hanning, dan Gaussian. Masing-masing window memiliki karakteristik tersendiri. Dalam penelitian ini digunakan metode Hamming Window. Berikut ini adalah persamaannya:

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N-1}\right), 0 \le n \le N \qquad (1)$$

dengan:
n = 0,1,....,(N-1)/2 , untuk N ganjil
n =0,1,.....,(N/2)-1 , untuk N genap

Hasil dari proses windowing ini adalah berupa suatu sinyal dengan persamaan:

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

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

3. Fast Fourier Transform

Tujuan utama dari transformasi fourier ini adalah untuk mengubah sinyal dari domain waktu menjadi spektrum pada domain frekuensi. Fast Fourier Transform (FFT) merupakan algoritma perhitungan Discrete Fourier Transform (DFT) yang efisien sehingga akan mempercepat proses perhitungan DFT. FFT dapat mereduksi jumlah perhitungan untuk setiap N data yang sama pada perhitungan DFT sehingga perhitungan yang ada menjadi lebih cepat, khususnya ketika nilai N yang digunakan cukup besar dengan mempergunakan persamaan:

$$Y(k) = \sum_{n=0}^{N-1} y(n) W_N^{nk} \qquad (3)$$

dengan:
k = 0,1,.....,N/2
$y(n)$ = sinyal masukan
$W_N = e^{-j2\pi/N} = \cos\left(\frac{2\pi}{N}\right) - J\,sin(\frac{2\pi}{N})$

4. Mel Frequency Wrapping

Skala mel-frekuensi adalah pemetaan frekuensi secara liniear untuk frekuensi di bawah 1 kHz dan

logaritmik untuk frekuensi di atas 1 kHz. Sebagai titik referensi, pitch dari 1 kHz, 40 dB diatas perceptual hearing threshold, didefinisikan sebagai 1000 mels. Oleh karena itu, dapat digunakan formula pada persamaan 2.4 berikut untuk menghitung mels frekuensi yang diberikan dalam Hz:

$$mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (4)$$

5. Cepstrum

Langkah terakhir dalam feature extraction yaitu mengubah kembali log mel spectrum ke dalam domain waktu. Hasilnya disebut mel frequency cepstrum coefficient (MFCC). Oleh karena itu, mel power spectrum coefficient tersebut merupakan hasil dari langkah terakhir yang dinotasikan dengan S_i, yang mana i = 1,2,....,M. Jadi MFCC, $\tilde{X}_k$ dapat dihitung dengan persamaan berikut.

$$\tilde{X}_k = \sum_{i=0}^{M-1} \alpha_k (\log S_i) \cos\left[\left(\frac{\pi((2i+1)k}{2M}\right)\right] \quad (5)$$

$$\propto_k = \begin{cases} \frac{1}{\sqrt{M}}, & k = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le k \le M-1 \end{cases}$$

dengan:
S_i = nilai *frequency wrapping*
k = *0, 2,.....,M-1*
M = jumlah *filter*

2.3. Kuantisasi Vektor

Kuantisasi vektor merupakan teknik kuantisasi klasik yang mana dilakukan pemodelan dari fungsi kepadatan probabilitas dengan distribusi vektor. Kuantisasi vektor mengelompokkan vektor ciri X={$\tilde{X}_k$ *: k* = 1, 2, ..., *K*} menjadi *codebook C= (c₁,c₂,.....,cₘ)*. Vektor x_k disebut sebagai vektor kode. Vektor-vektor ini merupakan vektor-vektor data yang diperoleh dari hasil ekstraksi yang disebut dengan *codeword.* Algoritma yang digunakan dalam pembentukan *codebook* adalah algoritma LBG (*Linde, Buzo, Gray)*. Algoritma LBG tersebut dapat diimplementasikan dengan prosedur rekursif sebagai berikut.

1. Mendesain vektor *codebook* yang merupakan centroid dari keseluruhan vektor *traning.*
2. Melipatgandakan ukuran dari *codebook* dengan membagi masing-masing *codebook* c_n menurut aturan:
   $c_n^+ = c_n(1+\varepsilon)$
   $c_n^- = c_n(1-\varepsilon)$
   yang mana n memiliki nilai dari satu sampai dengan *current size codebook* dan Ɛ adalah parameter *splitting* (Ɛ=0,0001).
3. Pencarian *Nearest-Neighbour*: mengelompokkan *training* vektor yang mengumpul pada blok tertentu. Selanjutnya menentukan *centroid* dalam *current codebook* yang terdekat dan memberikan tanda vektor yaitu *cell* yang diasosiasikan dengan *centroid-centroid* yang terdekat.
4. Pembaharuan *centroid:* menentukan *centroid* baru yang merupakan *codeword* yang baru pada masing-masing *cell* dengan menggunakan *training* vektor pada *cell* tersebut.
5. Iterasi I: mengulang langkah 3 dan 4 sampai diperoleh jarak penyimpangan rata rata (D) yang besarnya dibawah batasan yang telah ditentukan (δ). D' merupakan nilai distorsi awal yang nilainya ditentukan pada saat inisialisasi pada awal program.
6. Iterasi II: mengulang langkah 2,3 dan 4 sampai diperoleh *codebook* dengan ukuran M.

2.4. Perhitungan Jarak Penyimpangan

Perhitungan jarak penyimpangan dilakukan dengan membandingkan antara koefisien MFCC dari sinyal ucapan yang akan dikenali dan codebook dari tiap-tiap pengucap pada basisdata. Untuk menghitung jarak penyimpangan antara dua vektor maka digunakan jarak Euclidean dan jarak Mahalanobis (Mahalanobis distance). Jarak Euclidean didefinisikan dengan persamaan berikut.

$$d_E(x,y) = \sqrt{\sum_{i=1}^{dim}(x_i - y_i)^2} \quad (6)$$

Jarak Mahalanobis didefinisikan dengan persamaan berikut[13].

$$d(\vec{y},\vec{x}) = (y-\mu)^T \times \sum^{-1} \times (y-\mu) \quad (7)$$

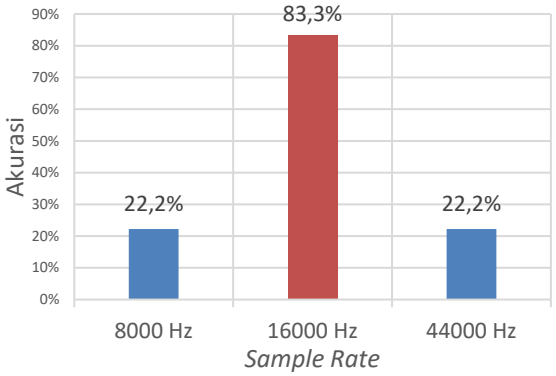
dengan:
$d_E(x,y)$ =jarak Euclidean
$d(x,y)$ =jarak Mahalanobis
y =nilai vektor pada suatu sampel suara yang di uji
x =nilai vektor pada suatu sampel suara pada *codebook*
μ =mean dari data sampel
T =transpose
$\sum^{-1}$ = invers covarian dari mean data sampel

3. Hasil dan Analisa

Hasil penelitian berasal dari 90 data suara responden yang dijadikan sebagai data pelatihan dan 36 data suara responden sebagai data pengujian. Pengujian yang dilakukan dibagi menjadi tiga tahap, yaitu:

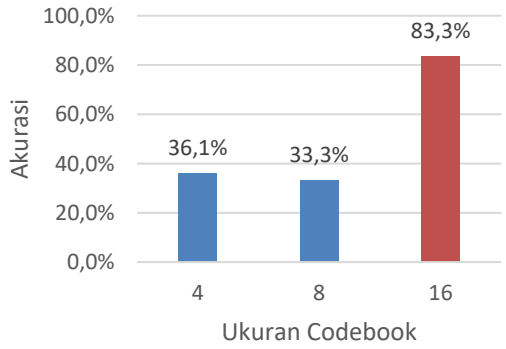
1. Pengujian dengan variasi sample rate yang berbeda-beda yaitu 8000Hz, 16000Hz, dan 44100Hz. Sample rate menunjukan nilai sinyal audio yang diambil dalam satu detik ketika melakukan rekaman suara. Semakin tinggi nilai dari sample rate kualitas audio yang direkam akan semakin baik. Berdasarkan grafik pada Gambar 4 dapat disimpulkan bahwa akurasi terbaik dicapai sistem dengan menggunakan sample rate 16000Hz. Hal ini disebabkan karena sinyal dengan sample rate 8000Hz belum dapat menyimpan

semua karakteristik suara yang diperlukan sehingga ciri yang diekstraksi belum sesuai dengan karakteristik yang diperlukan sistem, namun ketika sample rate diubah menjadi lebih tinggi (44100Hz) ternyata sistem tidak memberikan hasil yang lebih baik.



Gambar 5 Pengaruh nilai *Sample rate* terhadap akurasi.

2. Pengujian dengan variasi ukuran codebook, yaitu pengujian yang bertujuan untuk mengetahui ukuran codebook yang tepat untuk keberhasilan program yang sudah dibuat. Codebook merupakan representasi dari feature vector semua sinyal suara yang telah di clustering. Ukuran codebook dapat disesuaikan dengan kebutuhan sistem, namun pada Penelitian ini, ukuran codebook yang diobservasi adalah 4, 8, dan 16, untuk ukuran codebook >16 tidak dapat diterapkan pada sistem ini karena akan menghasilkan kegagalan pada proses perhitungan jarak Mahalanobis karena ketidaksesuaian jumlah matriks. Berikut adalah hasil pengujian akurasi sistem terhadap ukuran codebook.



Gambar 6 Grafik Akurasi sistem terhadap ukuran *Codebook.*

Pada Gambar 6 dapat dilihat bahwa akurasi tertinggi dicapai pada saat ukuran codebook berjumlah 16, sementara akurasi terendah didapat oleh codebook dengan ukuran 8. Dapat diamati bahwa semakin meningkatnya ukuran codebook maka akurasi sistem cenderung meningkat, selain itu jarak Mahalanobis yang didapat akan semakin besar pula mengikuti peningkatan ukuran codebook.

3. Pengujian dengan kondisi tidak ideal/salah, yaitu pengujian yang bertujuan untuk mengetahui keberhasilan program untuk mengenali data yang salah. Pada pengujian ini progam berhasil untuk tidak mengenali responden yang salah. Akurasi progam untuk tidak mengenali responden apabila diberikan suara responden yang salah sebesar 100%.

4. Kesimpulan

Berdasarkan hasil pengujian yang dilakukan , didapat beberapa point-point yang perlu diperhatikan diantaranya sample rate memberikan pengaruh yang signifikan yang mana didapat bahwa sample rate yang sesuai untuk program ini adalah sebesar 16000Hz, sedangkan pada sample rate 8000Hz dan 44100Hz hasil pengenalan yang didapat tidak begitu baik. Pada pengujian variasi ukuran codebook didapat hasil bahwa semakin tinggi ukuran codebook maka hasil yang didapat dalam sistem ini semakin baik, pada kasus ini ukuran codebook yang paling ideal adalah berjumlah 16, sedangkan pada ukuran codebook 4 dan 8 sistem belum mencapai hasil yang diinginkan. Pengujian terakhir dilakukan pada kondisi tidak ideal yang mana input yang dimasukkan pada saat pengujian salah. Hal ini dilakukan untuk menguji sistem bisa mengenali masukan yang salah atau tidak. Berdasarkan data hasil yang didapat, 100% data salah yang dijadikan masukan berhasil untuk tidak dikenali.

Referensi

[1] Ananda A, Ardha.2006. Penggunaan Pengenal Pengucap Tidak Berdasarkan Teks (*Speaker Recognition Text-Independent*) Sebagai Otorisasi Pengaksesan Pintu. Tugas Akhir S-1, Universitas Diponegoro, Semarang.

[2] Tiwari, Vibha. 2010 "*MFCC and Its Application in Speaker Recognition*". International Journal on Emerging Technologies, India.

[3] Gevaert, Wouter et al. 2010 "*Neural Networks used for Speech Recognition*". Journal of Automatic Control, Universitas Belgrade, Serbia.

[4] Buono, Agus. 2009. "Representasi Nilai HOS dan Model MFCC Sebagai Ekstraksi Ciri Pada Sistem Identifikasi Pembicara di Lingkungan Ber-Noise Menggunakan HMM". [disertasi]. Program Pascasarjana, Universitas Indonesia, Depok

[5] Jarwadi,(2008).*Speech To Text Menggunakan Database Diphone Dalam Bahasa Indonesia Dengan Metode Pendekatan Hybrid Hidden Markov Model Dan Algoritma Genetika.* Tugas Akhir S-1 , Universitas Telkom ,Bandung.

[6] Gold, Ben, and Nelson Morgan. Speech and Audio Signal Processing : *Processing and Perception of Speech and Music*, John Willey & Sons, Inc., New York,1999

[7] Apriyono, Fachrudin. (2009). *Pengenalan Pengucap Tak Bergantung Teks dengan Metode Vector Quantization (VQ) Melalui Ektraksi Linear Predictive Coding (LPC*). Tugas Akhir , Teknik Elektro, Universitas Diponegoro, Semarang.

[8] Campbell, Jr JP. 1997. *Speaker Recognition: A Tutorial Proceeding IEEE*. 85:1437- 1461.

[9] Hartaman, M. Rizky. (2009). *Rancang Bangun Sistem Pengenalan Penyakit Jantung dengan Metode Hidden Markov Model.* Tugas Akhir, Fakultas Teknik, Universitas Indonesia, Depok.

[10] Uchat, Nirav D. (2006). *Hidden Markov Model and Speech Recognition. Lecturer Handout.*

[11] Alfarisi, Lutfie Salman. (2007). *Speech Recognition dengan Hidden Markov Model menggunakan DSP Starter Kit TMS320C6713.* Tugas Akhir, Fakultas Teknik, Universitas Indonesia, Depok

[12] Ifeachor, Emmanuel C., Barrie W. Jervis. (2002). *Digital Signal Processing :Practical Approach. New Jersey : Prentice Hall.*

[13] W., Fawwaz Al Maki. (2000). *The Comparison of Vector Quantization Algorithms in Fish Species Acoustic Voice Recognition Using Hidden Markov Model.* Tugas Akhir , Fakultas Teknik, Universitas Indonesia, Depok.