Biospecies Vol. 9 No.1, Januari 2016, hal. 1-6.

Identifikasi *Speech Recognition* Manusia dengan Menggunakan *Average Energy* dan *Silent Ratio* Sebagai *Feature Extraction* Suara pada Komputer

Identification Human speech recognition using Average energy and Silent ratio as Voice feature extraction in Computer

Wawan KURNIAWAN[1)]

[1)]Staf Pengajar Pendidikan Fisika PMIPA FKIP Universitas Jambi
Kampus Pinang Masak Mendalo Darat Jambi
email: wwnkurnia79@gmail.com

Abstract. Biometric recognition is recognition systems or the identification of a person based on specific biological characteristics possessed by the person, the function of the security system by recognizing a person's identity. Speech recognition is a process performed computer to recognize words spoken by a person without regard to the identity of the person concerned. The pattern of increase and decrease sound signals are characteristic of a person in terms of speech, so that the average energy sound and silence duration threshold is the most important factor to detect a person's character and voice patterns. While processing algorithms Frequency Fourier Transform (FFT) is used to feature extraction components of the data in the frequency domain. The test results that indicate that a person's voice recorded with a duration of two seconds later the distribution of wave signal for 0.2 seconds Average amount of energy obtained was 2.323 and the Silent ratio of 0.249. Both of these values are the hallmark of someone who can be recognized by the computer.

Keywords: Avarege Energy, Fourier FrequencyTransform, Speech recognition, Silent Ratio

Abstrak. *Biometric recognition* merupakan sistem pengenalan atau identifikasi seseorang berdasarkan karakteristik biologis khusus yang dimiliki oleh orang tersebut, fungsinya untuk sistem keamanan dengan mengenali identitas seseorang. *Speech recognition* merupakan suatu proses yang dilakukan komputer untuk mengenali kata yang diucapkan oleh seseorang tanpa mempedulikan identitas orang terkait. Pola peningkatan dan penurunan signal suara merupakan ciri khas *(feature extraction)* dari seseorang dalam hal berbicara, sehingga antara energy rata-rata (*Average Energy*) bunyi dan lamanya ambang diam (*Silent Ratio*) merupakan faktor yang terpenting untuk mendeteksi karakter dan pola suara seseorang. Sedangkan proses algoritma *Fourier Frequency Transform* (FFT) digunakan untuk mencirikan komponen-komponen data pada domain frekuensi. Hasil pengujian yaitu menunjukkan bahwa suara seseorang yang direkam dengan durasi dua detik kemudian dilakukan pembagian sinyal gelombang selama 0,2 detik diperoleh besarnya *Average energy* adalah 2,323 dan proses *Silent ratio* sebesar 0,249. Kedua nilai tersebut merupakan *feature extraction* suara seseorang yang dapat dikenali oleh komputer.

Kata **kunci:** Avarege Energy, Fourier Frequency Transform, Speech recognition, Silent Ratio

PENDAHULUAN

Voice recognition dibagi menjadi dua jenis, yaitu *speech recognition* dan *speaker recognition*. *Speech recognition* adalah proses identifikasi suara berdasarkan kata yang diucapkan. Parameter yang dibandingkan ialah tingkat penekanan suara yang kemudian akan dicocokkan dengan *template database* yang tersedia. Sedangkan sistem pengenalan suara berdasarkan orang yang berbicara dinamakan *speaker recognition*. Pada penelitian ini hanya akan dibahas mengenai *speech recognition* karena kompleksitas algoritma yang diimplementasikan lebih sederhana dari pada *speaker recognition.*

Biometrik, termasuk di dalamnya *speech recognition*, secara umum digunakan untuk identifikasi dan verifikasi. Identifikasi ialah mengenali identitas seseorang, dilakukan perbandingan kecocokan antara data *biometric* seseorang dalam *database* berisi record karakter seseorang. Sedangkan verifikasi adalah menentukan apakah seseorang sesuai dengan apa yang dikatakan terhadap dirinya. *Biometric recognition* merupakan sistem pengenalan atau identifikasi seseorang berdasarkan karakteristik biologis khusus yang dimiliki oleh orang tersebut. Fungsinya selain untuk sistem keamanan dengan mengenali identitas seseorang, juga untuk identifikasi penyakit yang diderita seseorang, keperluan militer, dan lain-lain. Aplikasi *biometric recognition* antara lain *retinal scan* (identifikasi berdasarkan pola pembuluh darah pada retina mata), *fingerprint recognition* (identifikasi pola sidik jari unik pada setiap orang), *face recognition* (pengenalan seseorang berdasarkan raut dan ekspresi seseorang dengan kunci utama pada letak mata dan mulut), dan *voice recognition*.

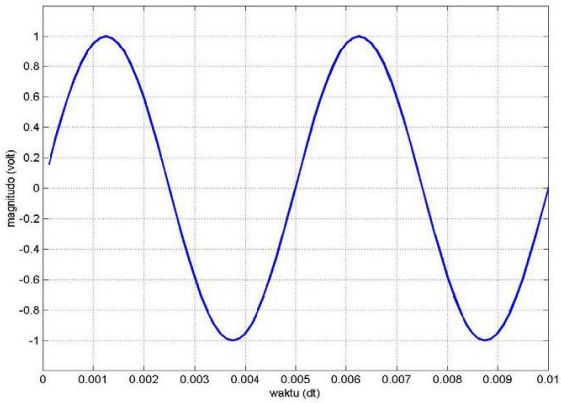
Voice recognition dibagi menjadi dua jenis, yaitu *speech recognition* dan *speaker recognition*. Berbeda dengan *speaker recognition* yang merupakan pengenalan identitas yang diklaim oleh seseorang dari suaranya (ciri khusus dapat berupa intonasi suara, tingkat kedalaman suara, dan sebagainya), *speech recognition* adalah proses yang dilakukan komputer untuk mengenali kata yang diucapkan oleh seseorang tanpa mempedulikan identitas orang terkait. Implementasi *speech recognition* misalnya perintah suara untuk menjalankan aplikasi komputer. Algoritma FFT (*Fourier Frequency transform*) merupakan salah satu metode untuk transformasi sinyal suara menjadi sinyal frekuensi. Artinya proses perekaman suara disimpan dalam bentuk digital berupa gelombang spektrum suara berbasis frekuensi.

Bentuk Sinyal dalam Gelombang Sinusoidal

Bentuk sinyal sinusoidal merupakan contoh sinyal waktu kontinyu dan juga menggunakan terminologi sinyal analog. Bentuk persamaan gelombang dalam persamaan (1) mempresentasikan nilai magnitudo sinyal sebagai fungsi waktu. Di dalam kondisi real seringkali dinyatakan dalam besaran volt. Nilai x(t) dalam parameter yang umum untuk pengukuran dinyatakan dalam V(t) yang penunjukkan nilai simpangan sinyal atau magnitudonya pada suatu waktu t.

$$x(t) = A\cos(2\pi t + \theta) \qquad (1)$$

bentuk sinyal sinusoidal seperti pada gambar di bawah ini.



Gambar 1. Grafik suara pola signal sinusoidal frekuensi 200 Hz

Sedangkan untuk besaran lain dari sinyal dalam hal ini adalah daya dinyatakan sebagai persamaan (2), sebagai perbandingan antara tegangan (V) dan hambatan (R):

$$P(t) = \frac{(V(t))^2}{R} \qquad (2)$$

Energi Pada Sinyal Bicara

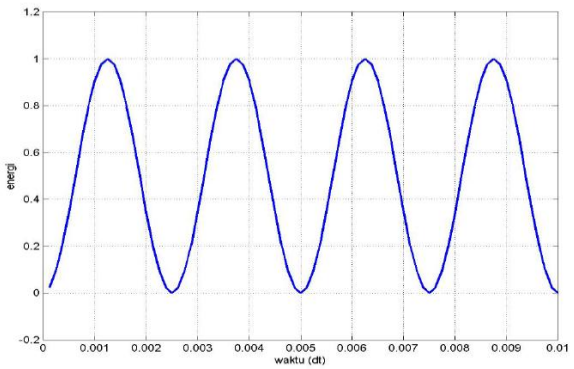
Untuk menghitung energi sinyal suara menggunakan formulasi dasar berikut:

$$E_k = \sum_{t=0}^{T} (V(t)w(t))^2 \qquad (3)$$

Dimana: w (m) = fungsi window seperti *hamming*, *honing*, *Bartlett*, dan *boxcar*. Panjang window dalam hal ini adalah m, untuk durasi dari t=0 sampai t=T akan didapatkan window sebanyak n=T/m apabila tidak ada *overlapping* antara window satu dengan yang lain. Jika terjadi *overlapping* antara window satu dengan yang lain, misalnya sebesar m/2, maka jumlah window dalam satu durasi T adalah sebanyak n = 1 + T/(m/2). Untuk suatu pengamatan energi pada *frame* ke-*k* bentuk persamaan (4) menjadi:

$$E_k = \sum_{t=0}^{T} (V(t)w(k-t))^2 \qquad (4)$$

dimana k akan menentukan posisi titik-titik window pada sinyal tersebut, ini juga dikenal sebagai model *sliding window*. Untuk sinyal sinus dalam bentuk energi dapat diberikan seperti gambar berikut:



Gambar 2. Sinyal sinus dalam bentuk energy

Silent Ratio (SR) adalah jumlah dari banyaknya sampel yang dibawah nilai treshold tertentu, dibagi dengan banyaknya sampel seperti rumus (5).

$$SR = \frac{\sum s}{\sum l} \qquad (5)$$

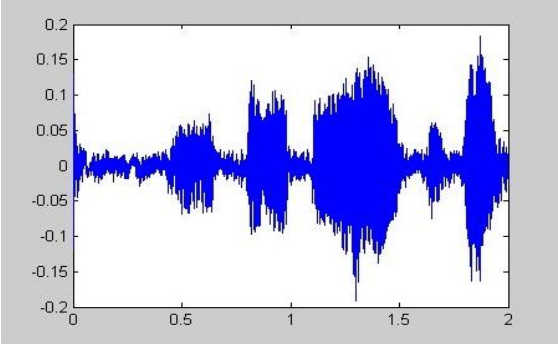
s = Periode keheningan dalam potongan file audio
l = Panjang dari tiap potongan file audio

Masukkan audio, apabila berpusat tinggi maka dikenal sebagai musik. Apabila tidak, maka perkataan dan musik. Apabila rasio keheningan tinggi, maka bukan musik, tetapi perkataan dan solo musik. Sehingga penelitian ini bertujuan untuk menyelidiki bagaiamana cara mengekstraksi ciri suara yang direkam sehingga dapat diketahui pola *average energy* dan *silent ratio* sebagai ciri dari suara manusia.

METODE PENELITIAN

1. Akuisisi Signal Audio

Pengambilan suara dilakukan melalui rekaman laptop selama 2 detik frekuensi 11.025 Hz.

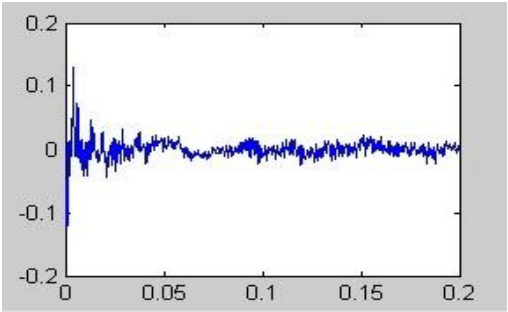


Gambar 3. Spektrum suara dalam 2 detik

2. Ekstraksi Ciri Suara

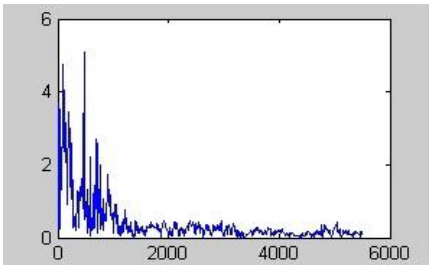
Untuk mengenali bentuk atau pola dari suatu sinyal suara harus dengan mengekstraksi ciri sinyal tersebut. Adapun langkah-langkah yang harus dilakukan adalah :

a. Membagi sinyal rekaman suara berdasarkan range waktu, sehingga diperoleh potongan-potongan sinyal suara dalam hal ini range waktu yang digunakan 0,2 detik.



Gambar 4. Pemotongan signal suara dalam 0,2 detik

b. Setiap potongan sinyal kemudian dirubah ke dalam domain frekuensi dengan menggunakan *Fourier Frequency Transform* (FFT)



Gambar 5. Bentuk domain frekuensi dengan FFT

c. Setelah dapat FFT maka dapat dilihat power atau energy tekanan tiap puncak sinyal terhadap frekuensi kemudian dihitung rata-rata energinya:

$$E = \frac{\sum_{n=0}^{N-1} x(n)^2}{N} \qquad (6)$$

Dimana: E = Energy
N = Banyaknya puncak signal
N = Nilai energy setiap signal

d. Ekstraksi ciri untuk suara kedua adalah *Silent ratio* dimana setiap akuisisi suara selalu ada suara yang berada dalam keadaan tenang (*noise*) hingga diambang diam. Hal ini perlu diperhitungkan karena untuk mendapatkan signal suara yang khas harus dikurangi dengan ambang keadaan diam (*silent).*

e. Setelah didapat *Avarage Energy* dan *Signal Ratio* setiap potongan kemudian dijumlahkan kesemua potongannya, sehingga dihasilkan nilai rata-rata untuk setiap rekaman suara

HASIL DAN PEMBAHASAN

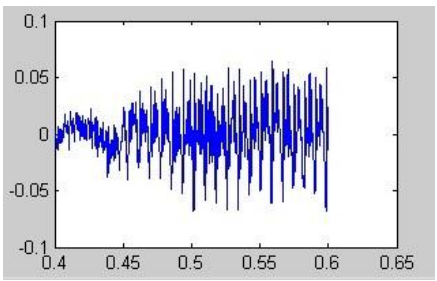
Hasil penelitian ini diperoleh nilai rata-rata Average Energy dan Silent Ratio, kemudian nilai tersebut digunakan sebagai ekstraksi suara yang akan dikenali oleh komputer. Hasil tersebut ditujukan pada Tabel 1.

Tabel 1. *Silent Ratio* dan *Avarage Energy* untuk Identifikasi Signal Suara

No.	Waktu	Silent Ratio	Average Energy
1	0,0 – 0,2	0,3985	0,5331
2	0,2 – 0,4	0,4316	0,1508
3	0,4 – 0,6	0,2135	1,0525
4	0,6 – 0,8	0,3636	0,552
5	0,8 -1,0	0,1061	4,1778
6	1,0 -1,2	0,2375	2,1395
7	1,2 – 1,4	0,0558	8,4383
8	1,4 – 1,6	0,2085	3,0165
9	1,6 – 1,8	0,2344	0,8545
	Jumlah	**2,2495**	**20,915**
	Rata -rata	**0,24994**	**2,3238**

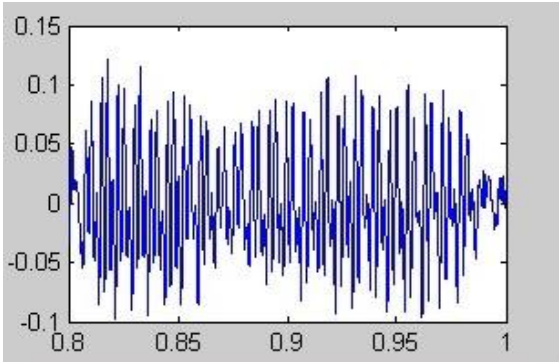
Sinyal suara yang diperoleh dari hasil pengambilan data suara selama 2 detik, dipotong atau dibagi dalam rentang 0,2 detik. Beberapa grafik hasil pemotongan menunjukan adanya penyimpangan amplitude yang besar sehingga menunjukan adanya energy lebih pada sinyal.

Pengambilan suara pada rentang waktu 0,4-0,6 yang menghasilkan energy sebesar 1,0525. Grafik hasil pemotongan suara untuk rentang waktu 0,4-0,6 dapat dilihat pada Gambar 6.



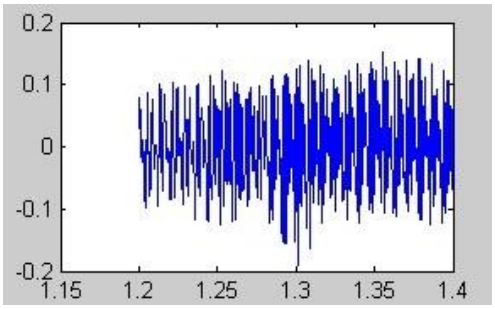
Gambar 6. Pemotongan sinyal suara pada 0,4-0,6 detik

Berikutnya pada rentang waktu 0,8-1,0 detik terjadi peningkatan energy akibat penyimpangan amplitudo dengan merata-ratakan setiap puncak pada sinyal suara, maka diperoleh energi suara tersebut adalah 4,1778, seperti pada Gambar 7.



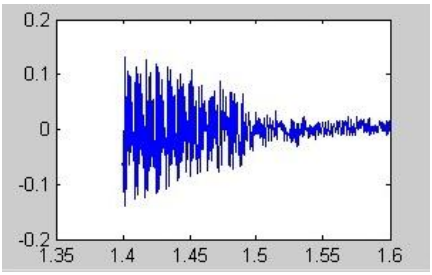
Gambar 7. Pemotongan sinyal suara pada 0,8-1,0 detik

Pada detik 1,2-1,4 kembali terdapat sinyal suara yang memiliki energi rata-rata 8,4383 dan penyimpangan amplitudo yang cukup besar, hal ini diakibatkan karena adanya intonasi dan intensitas suara yang kuat pada saat pembicara mengeluarkan bunyi sebuah kata. Grafik pada Gambar 8 mendiskripsikan peristiwa tersebut.



Gambar 8. Pemotongan sinyal suara pada 1,2-1,4 detik

Selanjutnya pada 1,4-1,6 detik terdapat peningkatan intensitas dan penyimpangan amplitudo sehingga menghasilkan energy rata-rata pada setiap sinyal adalah 3,0165, seperti pada Gambar 9.

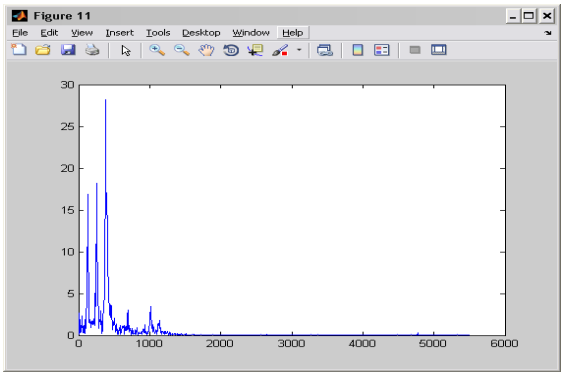


Gambar 9. Pemotongan sinyal suara pada 1,4-1,6 detik

Pola peningkatan dan penurunan signal suara merupakan ciri khas dari seseorang dalam hal berbicara, sehingga antara panjang bunyi dan lamanya ambang diam merupakan faktor yang terpenting untuk mendeteksi karakter dan pola suara seseorang (*speech recognition*).

Pengkuran nilai energi pada sinyal suara harus melibatkan fungsi window. Hal ini dikarena dalam pengukuran energi sinyal suara kita harus menyusunnya dalam frame-frame tertentu. Ini merupakan standar dalam teknologi speech processing, sebab secara umum dalam pengolahan sinyal suara kita terlibat dengan sinyal dengan durasi yang terlalu panjang bila dihitung dalam total waktu pengukuran. Fenomena ini juga dikenal sebagai *short term speech signal energy*.

Pada proses algoritma *Fourier Frequency Transform* (FFT), digunakan untuk mengestrak komponen-komponen data pada domain frekuensi, dimana domain spasial ataupun domain waktu, komponen-komponen tersebut tidak terlihat secara eksplisit. Seperti pada grafik yang dideskripsikan Gambar 10.



Gambar 10. Proses algoritma FFT dapat mengekstrak frakuensi

KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa bentuk khas suara (*feature extraction*) manusia dapat diperoleh dari besar energi rata-rata (*Average energy*) dan batas ambang diam (*Silent rasio*). Pola tersebut dapat dijadikan proses ekstraksi ciri pada proses pengenalan suara manusia. Data yang diperoleh dari penelitian ini bahwa besar *Average energy* adalah 2.323 dan proses *Silent ratio* sebesar 0.249.

DAFTAR PUSTAKA

Campbell, JPJR. 1997. *Speaker Recognition: A Tutorial.*Proc. IEEE, vol.85, no 9, pp1437-1462.

Proakis, JG, dan Manolakis, DG. 1997. *Pemrosesan Sinyal Digital*, edisi Bahasa Indonesia jilid 1, Jakarta:

Prenhallindo. Rabiner, LR, Juang BH. 1993. Fundamentals *of Speech Recognition*. New Jersey: Prentice Hall. ISBN 0-13-015157-2.

Ho, CE. 1998. *Speaker Recognition System*, Project Report. California: California Institut of Technology.

Krishnan, M, Neophytou, CP, and Prescott, G. 1994. *Wavelet Transform Speech Recognition Using Vector Quantization, Dynamic Time Warping and Artificial Neural Networks*, Lawrence, KS 66045:Center of excellence in computer aided systems engineering and Telecominication & Information Sciences Laboratory 2291 Irving Hill Drive

Li, Ze-Nian dan Drew, Mark S., 2004, Fundamentals of Multimedia, Pearson, Prentice Hall, Upper Saddle River, New Jersey,hal.137.