PENGENALAN SUARA MANUSIA DENGAN MENGGUNAKAN JARINGAN SARAF TIRUAN MODEL PROPAGASI BALIK

M. Ma'ruf Idris
Dosen Jurusan Teknik Elektronika Fakultas Teknik
Universitas Negeri Makassar

ABSTRAK

Pada penelitian ini dibuat sebuah sistem pengenalan suara manusia dengan jaringan saraf tiruan metode propagasi balik (back propagation) menggunakan personal komputer. Sinyal suara analog mula-mula dicuplik menjadi sinyal digital dengan kecepatan cuplik 8000 Hz. Untuk proses ekstraksi parameter suara digunakan metode Linear Predictive Coding (LPC) untuk mendapatkan koefisien cepstral. Koefisien cepstral LPC ini ditransformasikan ke dalam domain frekuensi dengan Fast Fourier Transform (FFT) 512 point. Hasil FFT selanjutnya diproses dengan jaringan saraf tiruan propagasi balik dengan konfigurasi neuron yaiut 32-160-100-30-30 untuk melakukan pengenalan. Lima puluh sampel suara dari lima pembicara yang berbeda digunakan sebagai input pada proses pelatihan jaringan saraf tiruan. Hasil pengujian proses pengenalan suara menunjukkan keberhasilan 90 %.

Kata Kunci : jaringan saraf tiruan propagasi balik, pengenalan suara, *LPC, FFT.*

I. PENDAHULUAN

Teknik jaringan saraf tiruan telah banyak dimanfaatkan pada berbagai bidang utamanya pada sistem pengenalan pola: citra, suara, *time series prediction*, dan lain-lain.

Pada penelitian ini dibuat suatu sistem memanfaatkan jaringan saraf tiruan metode propagasi balik (*back propagation)* untuk pengenalan suara. Sistem ini diharapkan dapat dimanfaatkan pada pemberian perintah komputer, *voice dialing*, dan lain-lain. Model jaringan saraf *back propagation* digunakan di sini bersama-sama dengan metode *linear predictive coding* dan *fast fourier transform* yang dipakai sebagai pemroses awal. Di sini dilakukan eksperimen variasi struktur dan parameter jaringan (jumlah *node hidden layer*, besarnya *step size* serta momentum) untuk mendapatkan *performance* jaringan yang optimum. Pencarian struktur dan parameter ini bertujuan agar jaringan dapat secara cepat belajar dan dapat mengenali suara dengan error sekecil mungkin.

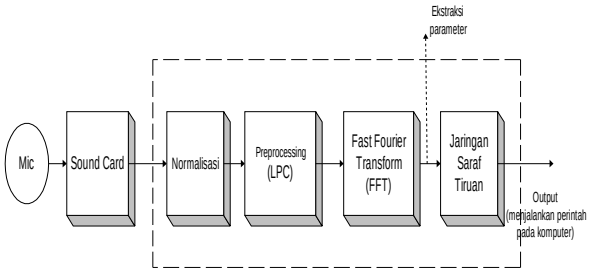
Lima puluh sampel suara dari lima pembicara yang berbeda digunakan sebagai input pada proses pelatihan jaringan saraf tiruan. Setelah dilakukan proses pelatihan,

sistem dicoba untuk mengenali suara pembicara. Hasil pengujian proses pengenalan suara menunjukkan keberhasilan sekitar 90 %.

2. Dasar Teori

2.1. SISTEM PENGENALAN SUARA

Sistem pengenalan suara yang dibuat digambarkan pada blok diagram gambar 1:



Gambar 1. Blok Diagram Sistem Pengenalan Suara Manusia

Secara garis besar, cara kerja sistem pengenalan suara ini ialah mula-mula sinyal suara manusia yang diterima dengan menggunakan *microphone* (sinyal *analog*) dicuplik sehingga menjadi sinyal *digital* dengan bantuan *sound card* pada PC.

Sinyal *digital* hasil cuplikan ini terlebih dulu dinormalisasi kemudian diproses awal menggunakan metode *LPC* sehingga didapat beberapa koefisien *LPC* yang merupakan feature (ciri) dari suara pembicaraan. Kemudian koefisien *LPC* tersebut diproses dengan *Fast Fourier Transform* (FFT) untuk mendapatkan sinyal pada domain frekuensi. Hal ini bertujuan agar perbedaan antar pola kata yang satu dengan yang lain terlihat lebih jelas sehingga ekstraksi parameter sinyal memberikan hasil yang lebih baik. Hasil keluaran *FFT* ini merupakan masukan bagi jaringan saraf tiruan *Back Propagation* dimana jaringan saraf tiruan ini berfungsi sebagai utama dari sistem untuk proses pengenalan suara.

2.2. Proses Pencuplikan Sinyal

Pencuplikan dilakukan pada kecepatan 8000 Hz dengan resolusi 8 bit (1 byte). Kecepatan pencuplikan tersebut dilakukan dengan didasarkan asumsi bahwa sinyal percakapan (*speech*) berada pada daerah frekuensi 300-3400 Hz sehingga memenuhi kriteria Nyquist yang menyatakan :

$$f_s \geq 2xf_h \quad f_h = f_{in}tertinggi \qquad \textbf{(1)}$$

Pada sinyal yang didapat kemudian dilakukan proses normalisasi dengan tujuan mendapatkan sinyal dengan ukuran yang sama walaupun kata yang diucapkan berbeda. Cara kerja proses normalisasi ini dilakukan dengan menambahkan beberapa data tambahan apabila data hasil pencuplikan belum memenuhi jumlah yang dibutuhkan atau dengan mengurangi jumlah data hasil pencuplikan apabila melebihi jumlah input yang dibutuhkan. Jumlah data output dari proses normalisasi ini ditetapkan sebanyak 3360 buah (0,42 detik) dengan asumsi bahwa untuk

pengucapan satu kata dibutuhkan waktu kurang dari 0,5 detik.

2.3. Preprocessing Sinyal Dengan Metode Linear Predictive Coding

Langkah-langkah analisa *LPC* untuk pengenalan suara adalah sebagai berikut:

- ***Preemphasis.*** Pada langkah ini, cuplikan kata dalam bentuk *digital* ditapis dengan menggunakan *FIR filter* orde satu untuk meratakan spektral sinyal kata yang telah dicuplik tersebut.

$$\tilde{s}(n) = s(n) - \tilde{\alpha} s(n-1) \qquad \textbf{(2)}$$

- ***Frame Blocking.*** Pada tahap ini sinyal kata yang telah teremphasi dibagi menjadi frame-frame dengan masing-masing frame memuat N cuplikan kata dan frame-frame yang berdekatan dipisahkan sejauh M cuplikan.
- ***Windowing.*** Pada langkah ini dilakukan fungsi *weighting* pada setiap frame yang telah dibentuk pada langkah sebelumnya.

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

$$0 \le n \le N-1 \qquad \textbf{(3)}$$

- ***Analisa Autokorelasi.*** Pada tahap ini masing-masing frame yang telah di *windowing* diautokorelasikan dengan nilai autokorelasi yang tertinggi adalah orde dari analisa LPC, biasanya orde LPC tersebut 8 sampai 16.

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \qquad \textbf{(4)}$$

- ***Analisa LPC.*** Langkah selanjutnya adalah analisa LPC, dimana pada tahap ini nilai autokorelasi pada setiap frame diubah menjadi satu set LPC parameter yaitu koefisien LPC, koefisien pantulan (*reflection coefficient*), dan koefisien perbandingan daerah logaritmis (*log area ratio coefficient*).
- ***Mengubah LPC Parameter ke Koefisien Cepstral.*** Koefisien cepstral ini merupakan koefisien transformasi Fourier yang merepresentasikan spektrum *log magnitude*.

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \;,\; 1 \le m \le p$$

(5)

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k},\; m > p \qquad \textbf{(6)}$$

Proses *frame blocking* yang dilakukan pada sistem ini ditetapkan tiap 30 mili detik dengan jarak antar *frame* 10 mili detik. Jadi dengan kecepatan cuplik sebesar 8000 Hz maka tiap *frame* akan berisi 240 byte data dengan jarak antar *frame* 80 byte data atau dengan kata lain *overlap* yang terbentuk sebesar 160 byte data. Dengan ketentuan *frame* seperti di atas, maka untuk data hasil cuplik sebanyak 3360 data maka akan terbentuk $\frac{3360-160}{80} = 40$ buah frame.

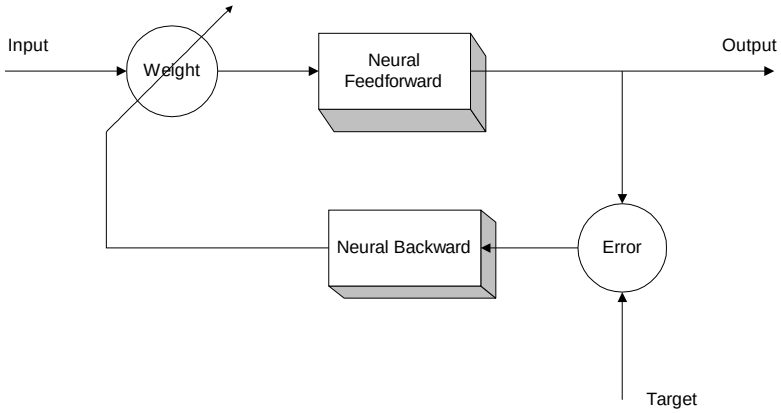
Untuk perhitungan koefisien cepstral, digunakan orde LPC 12 sehingga didapat data output sebanyak 40x12 = 480 data.

2.4. Fast Fourier Transform (FFT)

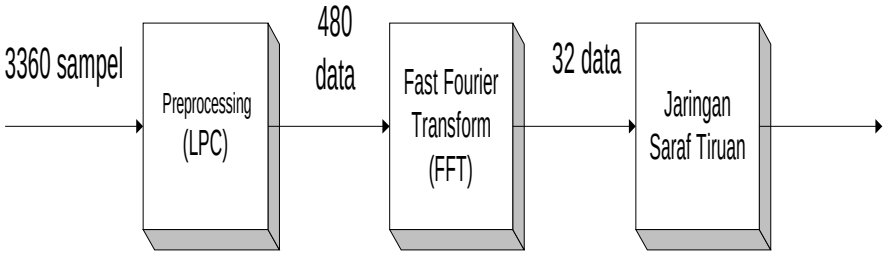
Proses *Fast Fourier Transform (FFT)* ini dilakukan setelah didapat koefisien cepstral sebanyak 480 data. FFT yang digunakan memakai 512 *point* dan karena hasil FFT simetris maka keluaran FFT tersebut hanya diambil sebanyak 256 data. Dari 256 data ini kemudian dibagi menjadi 32 blok dimana masing-masing blok berisi 8 data dan dihitung rata-rata untuk masing-masing blok. Maka total keluaran dari *FFT* ini adalah 32 data, dimana data tersebut merupakan masukan bagi jaringan saraf tiruan.

2.5. Jaringan Saraf Tiruan Back Propagation

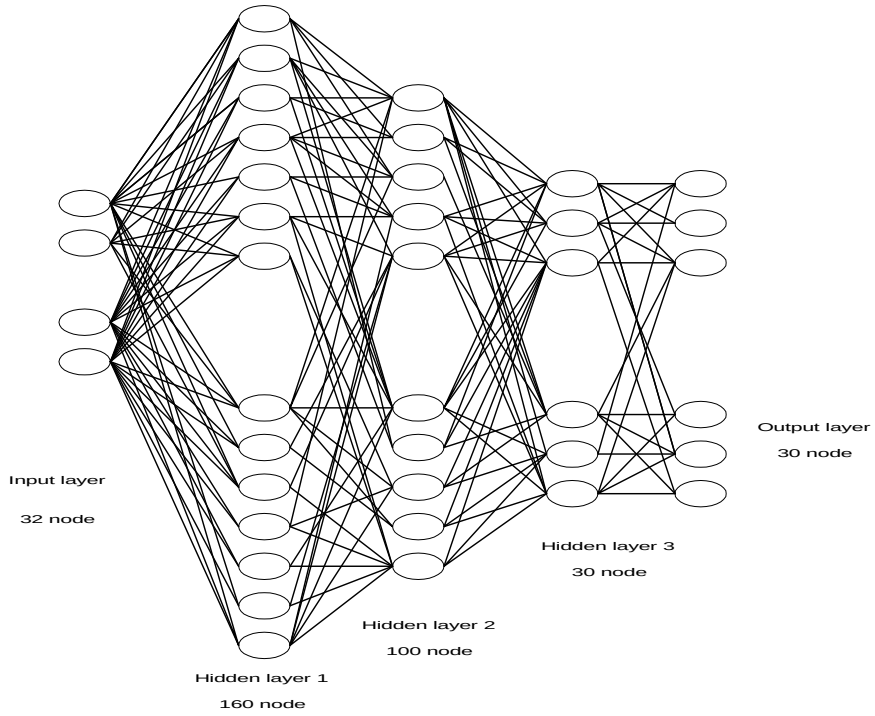
Pada sistem ini, input jaringan saraf tiruan berasal dari keluaran *FFT* yang telah dibagi menjadi 32 blok. Jadi terdapat 32 data input bagi jaringan saraf tiruan. Sedangkan outputnya berjumlah 30 buah yang masing-masing merupakan bilangan biner dan masing-masing bit merepresentasikan satu buah pola kata. Jadi pada output terdapat 30 kemungkinan yang dapat dihasilkan.



Gambar 2. Blok Diagram Jaringan Syaraf Tiruan Propagasi balik (*Backpropagation)*



Gambar 3. Struktur *LPC* dan *FFT*.



Gambar 4. Struktur Jaringan Syaraf Tiruan.

Pelatihan jaringan dilakukan dengan mengambil input dari pembicara sebanyak 5 orang (pembicara 1, pembicara 2, pembicara 3, pembicara 4 dan pembicara 5) dimana masing-masing pembicara mengucapkan 10 buah pola kata yaitu ‘nol’, ‘satu’, ‘dua’ dan seterusnya sampai pola kata ‘sembilan’. Pola kata dari masing-masing pembicara tersebut disimpan dan kemudian dilatihkan secara bersamaan ke dalam jaringan saraf tiruan. Pola kata tersebut dimasukkan secara urut mulai pembicara 1 dengan pola kata ‘nol’, ‘satu’, ‘dua’ dan seterusnya sampai pola kata ‘sembilan’, kemudian pembicara 2 dengan pola kata ‘nol’, ‘satu’, ‘dua’ sampai pola kata ‘sembilan’, demikian seterusnya sampai pembicara 5. Setelah semua data dimasukkan maka proses training dilakukan sampai error yang dihasilkan mencapai nilai yang telah ditentukan dimana pada proses ini digunakan nilai error 0,0001.

3. HASIL-HASIL PERCOBAAN

3.1.Penentuan Struktur Jaringan Saraf Tiruan

Pertama kali akan ditentukan jumlah *node hidden layer* dengan menggunakan satu *hidden layer*, nilai *learning rate* 0,5 momentum 0,5 serta input 32 data. Dari hasil pengujian didapat bahwa makin banyak jumlah *node hidden layer* yang digunakan maka akan menghasilkan error yang kecil dalam iterasi yang makin singkat, sampai

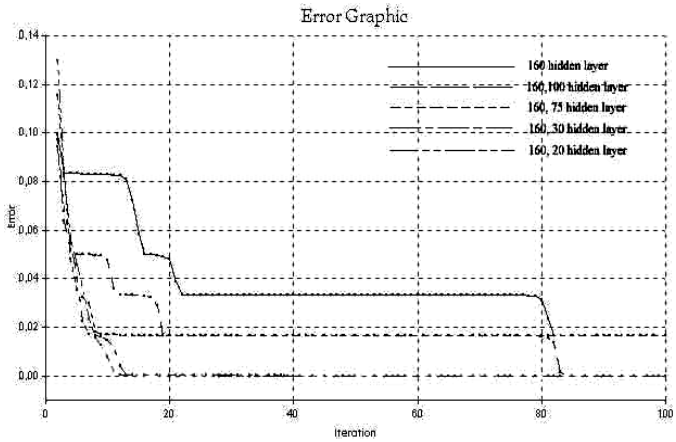
mencapai suatu nilai tertentu dimana perubahan jumlah *node* hanya mengakibatkan sedikit perubahan pada jumlah iterasi. Dari pengujian tersebut maka didapat bahwa jumlah *node hidden layer* yang optimal ialah 160 buah. Penambahan jumlah *node* lebih besar dari 160 tidak menghasilkan penurunan jumlah iterasi yang berarti. Makin banyak jumlah *node* yang digunakan akan memakai memori komputer makin besar sehingga jika dipilih jumlah *node* diatas 160 akan terdapat pengorbanan pada jumlah memori yang digunakan tanpa diiringi perubahan jumlah iterasi yang berarti.

Setelah diketahui jumlah *node hidden layer* yang optimum, maka kemudian dilanjutkan untuk menentukan nilai *learning rate* yang optimum. Untuk itu jaringan saraf tiruan akan diuji dengan menggunakan 1 hidden layer dengan 160 *node*, serta momentum 0,9, 0,75, 0,5, 0,25 dan 0,1. Dari pengujian dapat diambil kesimpulan bahwa makin besar nilai *learning rate* yang digunakan, maka jumlah iterasi yang dibutuhkan untuk mencapai error yang kecil makin sedikit. Tetapi penggunaan nilai *learning rate* yang terlalu besar akan memperbesar kemungkinan error yang terjadi. Sehingga nilai *learning rate* yang baik tercapai pada nilai yang tidak terlalu besar ataupun terlalu kecil. Dari hasil eksperimen, diambil nilai *learning rate* 0,5 sebagai nilai yang terbaik.

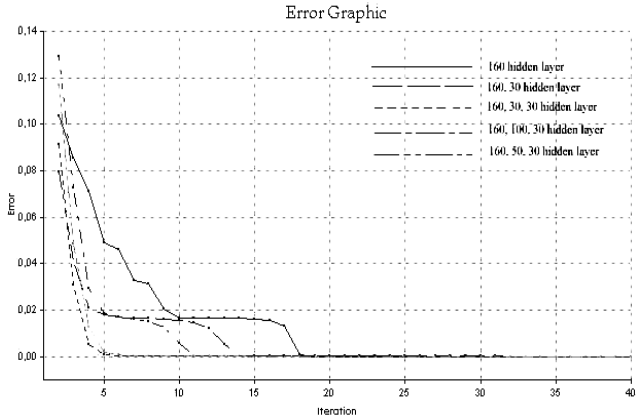
Selain menentukan nilai *learning rate*, maka perlu ditentukan pula nilai momentum yang optimum. Dari pengujian yang dilakukan didapat bahwa makin kecil nilai momentum maka makin banyak iterasi yang dibutuhkan untuk mencapai error yang kecil. Untuk itu perlu diambil nilai momentum yang optimum dimana dalam program ini diambil nilai momentum 0,75. Hal ini disebabkan karena dengan menggunakan momentum 0,75 akan diperoleh error yang kecil dengan jumlah iterasi yang tidak terlalu banyak ataupun terlalu sedikit.

Setelah menentukan parameter-parameter untuk satu *hidden layer*, maka sekarang akan diuji respon sistem jika menggunakan *hidden layer* lebih dari satu, dengan menggunakan nilai *learning rate* dan momentum yang didapat dari pengujian di atas. Dari pengujian ini (gambar 4 dan 5) didapat bahwa sistem akan optimal jika memakai struktur tiga hidden layer dengan konfigurasi 160, 100 dan 30, dimana dengan struktur tersebut dicapai error yang kecil dalam iterasi yang singkat. Gambar 6 menunjukan jika digunakan jumlah *hidden layer* lebih dari 3 maka akan didapat jumlah iterasi yang lebih banyak. Jadi dari semua pengujian yang dilakukan didapat struktur jaringan

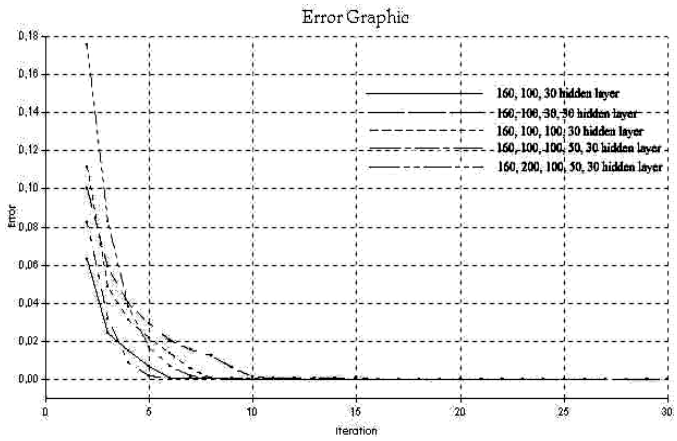
saraf tiruan yang optimal untuk sistem pengenalan suara ini ialah menggunakan 3 *hidden layer* dengan jumlah *node* 160, 100 dan 30 serta nilai *learning rate* yang digunakan 0,5 dan momentum 0,75.



Gambar 5. Grafik error dengan satu dan dua *hidden layer*



Gambar 6. Grafik error dengan satu, dua dan tiga *hidden layer*.



Gambar 7. Grafik error dengan tiga, empat dan lima hidden layer

3.2.Uji Pengenalan Suara

Pada tahap awal uji pengenalan dilakukan terhadap sinyal suara yang sama persis dengan yang telah ditrainingkan (*training data set*) dan didapat hasil bahwa error yang terjadi sebesar 2 % atau dengan kata lain keakuratan sistem untuk mengenali pola *training data set* mencapai 98 % (Tabel 1).

Kemudian dilakukan pengujian terhadap sinyal suara secara langsung dari *microphone* oleh orang yang sama dengan yang telah dilatihkan (pembicara1 s/d pembicara 5) ataupun oleh orang yang belum pernah dilatihkan sebelumnya yaitu pembicara 6 s/d 8 (*blind data set*). Dari proses pengujian ini didapat error rata-rata sebesar 10 % atau dengan kata lain keakuratan sistem untuk pengenalan pola *blind data set* mencapai 90 % (Tabel 2). Di sini tampak bahwa untuk pengenalan kata pada pembicara 1 s/d 5 terdapat error yang rendah, sedangkan pada pembicara 6 s/d 8 tampak error lebih tinggi, namun masih tetap bisa dikenali dengan kesalahan sekitar 20 %.

Tabel 1. *Error Rate* pada pengujian dengan *Training Data Set*

PEMBICARA	ERROR RATE
Pembicara 1	0%
Pembicara 2	0%
Pembicara 3	0%
Pembicara 4	**10%**
Pembicara 5	0%
Error rata-rata	**2%**

Tabel 2. *Error Rate* pada pengujian dengan *Blind Data Set*

PEMBICARA	ERROR RATE
Pembicara 1	0 %
Pembicara 2	0 %
Pembicara 3	10 %
Pembicara 4	10 %
Pembicara 5	0 %
Pembicara 6	20 %
Pembicara 7	20 %
Pembicara 8	20 %
Error rata-rata	**10 %**

4. KESIMPULAN

Berdasarkan hasil eksperimen yang telah dilakukan ini, maka dapat disimpulkan beberapa hal sebagai berikut di bawah ini:

1. Struktur jaringan saraf tiruan yang optimal untuk sistem pengenalan suara ini menggunakan 3 *hidden layer* dimana masing-masing *node* adalah 160, 100 dan 30 buah. Nilai *learning rate* yang digunakan 0,5 dan momentumnya 0,75. Keakuratan sistem pengenalan suara untuk pengenalan *training data set* mencapai 98 % dan untuk pengenalan *blind data set* mencapai 90 %.
2. Kesalahan pengenalan yang terjadi diakibatkan adanya perbedaan yang

terlalu besar antara sinyal suara yang hendak dikenali dengan sinyal suara yang dilatihkan, hal ini dapat diatasi dengan menambahkan/memperbanyak berbagai variasi pola kata pada saat pelatihan dengan demikian sistem jaringan lebih diperkaya pengetahuannya.

3. Terbuka penelitian lanjutan untuk memperbesar jumlah perbendaharaan kata, dan penggunaan metode *hibrid* lainnya sehingga pengenalan kata bersifat *speaker independent.*

DAFTAR PUSTAKA

[1] Freeman, James.A. *Neural Network Algorithms, Applications, and Programming Techniques*, Addison-Wesley Publishing Company, Inc., 1991.

[2] Fausett, Laurene. *Fundamentals Of Neural Network.* Englewood Cliffs, New Jersey : Prentice-Hall.Inc., 1994.

[3] Rabiner, L.R., Juang, B.H. *Fundamentals Of Speech Recognition.* Englewood Cliffs, New Jersey : Prentice-Hall.Inc., 1993.

[4] Oppenheim, Alan.V. *Discrete-Time Signal Processing*. Englewood Cliffs, New Jersey : Prentice-Hall.Inc., 1989.

[5] Orfanidis, Sophocles.J. *Optimum Signal Processing.* Singapore : McGraw-Hill Book Co., 1990.

[6] Eberhart, Russell.C. *Neural Network PC Tools.* San Diego, California : Academic Press. Inc., 1990.

[7] Todd, Bill., Kellen, Vince. *Delphi A Developer's Guide*. New York : M&T Books, 1995.