Pengenalan Ucapan Suku Kata Bahasa Lisan Menggunakan Ciri LPC, MFCC, dan JST

Abriyono*[1], Agus Harjoko[2]
[1] Jurusan Teknik Informatika, STMIK Widya Dharma, Pontianak
[2]Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
e-mail: ***[1]blacks_giants@yahoo.com**, [2]aharjoko@ugm.ac.id

Abstrak

Suara adalah salah satu alat komunikasi antar manusia yang efektif dan digemari. Selain sebagai alat komunikasi antar manusia, suara manusia telah digunakan sebagai alat komunikasi antara manusia dan komputer (mesin). Penelitian menggunakan suara sebagai alat komunikasi manusia dan mesin telah banyak dilakukan dengan menggunakan berbagai bahasa. Bahkan ada beberapa penelitian yang telah menghasilkan kemampuan pengenalan yang baik dan dikomersilkan (menggunakan bahasa Inggris). Bagaimana dengan penelitian pengenalan suara menggunakan Bahasa Indonesia? Peneliti mengamati penelitian pengenalan ucapan kata dalam Bahasa Indonesia masih minim dan cakupan jumlah katanya pun masih kecil. Oleh karena itu, pada penelitian ini, peneliti melakukan pengenalan ucapan kata Bahasa Indonesia. Pengenalan ucapan kata Bahasa Indonesia dilakukan dengan memecah kata Bahasa Indonesia ke dalam bentuk suku kata bahasa lisan. Pemecahan ke dalam bentuk lafal kata diharapkan mampu mengurangi jumlah kata yang sangat besar, namun tetap mengakomodasi seluruh kata yang dalam Bahasa Indonesia. Total jumlah lafal kata yang ditemukan oleh peneliti adalah 1741 suku kata bahasa lisan. Peneliti membagi sistem dalam 4 bagian besar, yakni proses perekaman, pre-processing, ekstraksi ciri, dan pengenalan. Pada proses perekaman digunakan frekuensi 11025 Hz, Mono, 8 bit. Pada pre-processing digunakan proses bantuan seperti pre-emphasis, segmentasi, framing, dan windowing. Sedangkan untuk ekstraksi ciri dan pengenalan digunakan ciri LPC/MFCC dan identifier jaringan syaraf tiruan backpropagation. Hasil pengenalan dengan pendekatan yang dibangun menunjukkan hasil yang belum memuaskan, yakni dengan kemampuan pengenalan terbaik sebesar 0.65% dengan ciri MFCC.

Kata kunci*—pengenalan kata Bahasa Indonesia, LPC, MFCC, JST, backpropagation.*

Abstract

Voice is one of effective and convinienced communication's medium among human. However, the used of voice is not only for communication among human but also has another role nowadays. Voice becomes communication medium for human and computer (machine). One of its application is speech to text application. Some of speech to text research already claimed good accuracy for some languages. How about Indonesian language? The research for Indonesian word recognition was still at low amount. The word used for research was at small amount too. Because of some of the reasons, researcher focus on Indonesian word recognition in this research. This research will divide the word into the speech syllable. The aim for the dividing system is to reduce the large amount of the word, but still cover all of the word. We found and used 1741 speech syllables. For managing the recognition, we used several approaches. The approaches are 11025 Hz, Mono, 8 bit for recording, pre-emphasized, segmentation, framing, and windowing for pre-processing, LPC and MFCC for the features, and back-propagation neural network for the identifier. The result using this approach was not reached good performance. The best result performed 0.65% by using MFCC feature.

Keywords*—Indonesian's syllable recognition, LPC, MFCC, neural network, backpropagation*

Received March 5th,2012; Revised June 20th, 2012; Accepted June 29 th, 2012

1. PENDAHULUAN

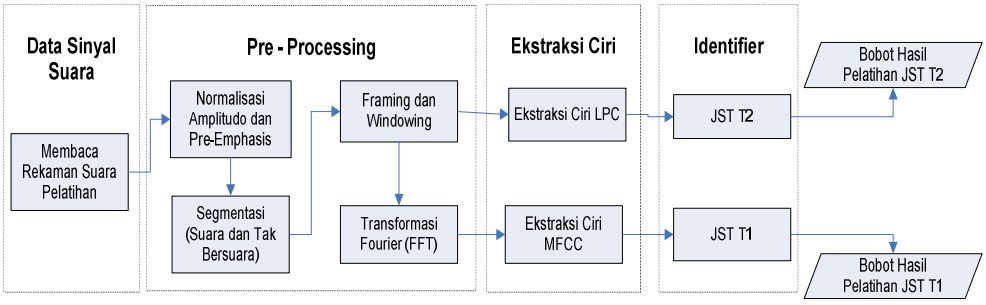
Suara adalah salah satu alat komunikasi yang digemari dan ‘efektif’ di antara manusia. Selain sebagai alat komunikasi antar manusia, suara manusia telah digunakan pada taraf yang lebih besar saat sekarang. Suara telah dapat dijadikan sebagai alat komunikasi antara manusia dan mesin (komputer). Salah satu aplikasi yang menggunakan suara adalah aplikasi speech to text. Penelitian speech to text telah banyak dilakukan dalam berbagai bahasa dunia dan telah ada yang mengklaim mencapai kemampuan yang baik serta dikomersilkan (Bahasa Inggris).

Bagaimana dengan pengenalan yang menggunakan Bahasa Indonesia? Sejauh penulis membaca beberapa penelitian, penelitian dengan menggunakan Bahasa Indonesia sebagai induk bahasa masih dalam jumlah yang terbatas. Penelitian penggunaan Bahasa Indonesia pun terbatas pada beberapa jumlah target kata seperti pada [1], [2], dan [3] serta hanya berfungsi untuk perintah suatu aplikasi tertentu seperti pada [4] dan [5]. Apakah yang menjadi kendala dalam pengenalan suara ucapan kata Bahasa Indonesia? Apakah Bahasa Indonesia terlalu sulit untuk dikenali? Kedua pertanyaan inilah yang muncul di benak penulis dan mendorong penulis untuk melakukan penelitian tentang pengenalan suara ucapan kata Bahasa Indonesia secara menyeluruh dan mencari tahu permasalahan yang terdapat didalamnya.

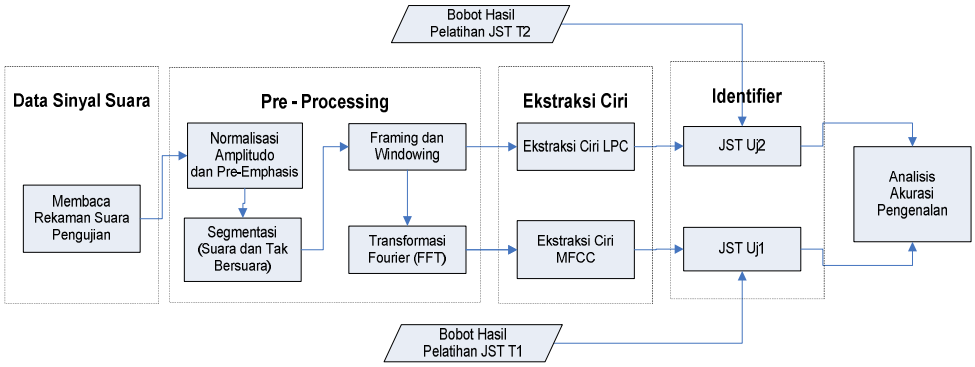
Pada penelitian ini, penulis lebih memfokuskan terlebih dahulu kepada penemuan teknik pengolahan dan pengenalan sinyal suara yang baik, sehingga penulis melakukan pembatasan pada gaya pengucapan suara yang seperti proses mendikte kata kepada anak yang baru belajar menulis, pembatasan tidak adanya gangguan latar dengan melakukan perekaman pada kondisi tenang, pembatasan karakteristik mic dengan menggunakan Mic Sennheiser PC110, dan menggunakan Bahasa Indonesia sebagai induk bahasa.

2. METODE PENELITIAN

Pada penelitian ini, sistem pengenalan dibagi ke dalam empat proses utama, yakni proses perekaman data sinyal suara, proses pre-processing, proses ekstraksi ciri, dan proses identifikasi. Proses perekaman suara adalah proses melakukan perekaman suara manusia yang mengucapkan kata dalam Bahasa Indonesia. Proses perekaman dilakukan pada kondisi tenang (sinyal dalam keadaan satu garis lurus dan bernilai mendekati nol). Proses pre-processing adalah proses untuk melakukan pengolahan awal terhadap data rekaman dengan mengurangi efek dari perekaman dan melakukan penguatan sinyal digital hasil perekaman, serta mempersiapkan data pada bentuk yang tepat untuk proses ekstraksi ciri. Proses pre-processing yang terlibat adalah proses normalisasi amplitudo [6] dan pre-emphasis [7], proses segmentasi [8], proses framing, proses windowing [9], dan proses fourier transform [10]. Proses ekstraksi ciri adalah proses mengambil nilai tertentu dari sinyal digital yang dapat mewakili sinyal digital secara keseluruhan. Pengambilan ciri yang tepat akan dapat membedakan antara suara yang satu dengan yang lainnya dan akhirnya meningkatkan akurasi pengenalan. Proses ekstraksi ciri yang dilakukan pada penelitian ini adalah ekstraksi ciri LPC [7] dan MFCC [11]. Proses terakhir, proses identifikasi, adalah proses untuk mengenali suatu sinyal suara. Proses identifikasi yang digunakan adalah jaringan syaraf tiruan model backpropagation [12].



Gambar 1 Bagan Alir Sistem Pengenalan Skema Pelatihan



Gambar 2 Bagan Alir Sistem Pengenalan Skema Pengujian

Dikarenakan proses identifikasi menggunakan jaringan syaraf tiruan yang memerlukan proses pelatihan terlebih dahulu, maka proses-proses tersebut dibagi ke dalam dua skema besar, yakni skema pelatihan dan skema pengujian. Adapun urut-urutan proses kedua skema ini ditunjukkan pada Gambar 1 dan 2. Gambar 1 menunjukkan tahapan proses untuk skema pelatihan dan Gambar 2 menunjukkan tahapan proses untuk skema pengujian. Proses-proses yang ada pada kedua skema ini hampir sama. Perbedaan utama pada kedua skema ini adalah pada data yang digunakan dan adanya proses analisis akurasi hasil pada skema pengujian. Hasil dari skema pelatihan yang berupa bobot jst digunakan kembali pada skema pengujian. Pada kedua skema bagian identifier, terdapat dua kotak proses untuk jaringan syaraf tiruan (T1 dan T2). Penggambaran ini menyiratkan proses perbandingan kemampuan pengenalan dari kedua ciri oleh penulis. Kedua ciri ini akan diujikan secara terpisah. Jaringan syaraf tiruan dengan kode 1 menggunakan masukan berupa ciri MFCC dan jaringan syaraf tiruan dengan kode 2 menggunakan masukan berupa ciri LPC.

2.1 Data Sinyal Digital (Rekaman Suara)

Data yang digunakan adalah data rekaman berisi suara ucapan kata Bahasa Indonesia. Kata Bahasa Indonesia yang digunakan tidaklah murni kata Bahasa Indonesia. Penulis menggunakan suku kata bahasa lisan Bahasa Indonesia [13]. Suku kata bahasa lisan adalah penggalan kata Bahasa Indonesia berdasarkan pemberhentian pada lafal (cara ucap) kata yang dilakukan. Perlu diingat bahwa suku kata bahasa lisan berbeda dengan suku kata bahasa tulis. Sebagai contoh kata ‘belajar’ yang memiliki suku kata bahasa lisan be, la, dan jar, sedangkan suku kata bahasa tulisan-nya bel-a-jar. Penggunaan suku kata bahasa lisan Bahasa Indonesia dilakukan untuk mengurangi jumlah data yang harus dikenali tanpa mengurangi jumlah kata Bahasa Indonesia yang ada. Penulis menemukan dan menggunakan 1741 suku kata bahasa lisan dalam Bahasa Indonesia.

Pada penelitian ini, perekaman data dilakukan pada frekuensi sampling (Fs) 11025 Hz, mono, 8 bit dan direkam pada kondisi ruangan tenang. Kondisi lain yang berhubungan dengan data adalah sumber suara. Sumber suara akan diambil dari tiga orang yang di mana setiap orang mengucapkan setiap suku kata bahasa lisan sebanyak tiga (3) kali. Dari tiga kali ucapan tersebut, dua rekaman pengucapan suku kata dari setiap orangnya akan dimasukkan sebagai data rekaman suara kelompok pelatihan sedangkan sisanya dimasukkan sebagai data rekaman suara kelompok pengujian.

2.2 Normalisasi Amplitudo

Normalisasi amplitudo adalah proses yang digunakan untuk menormalkan degradasi nilai sampel sinyal digital yang diakibatkan perbedaan jarak antara mulut dan mikrofon perekam. Proses normalisasi amplitudo diperoleh dengan membagi semua nilai sampel sinyal digital dengan nilai absolut maksimum dari sampel sinyal digital tersebut (lihat persamaan (1)).

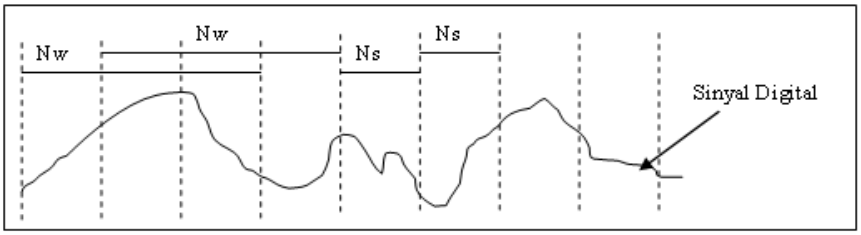
$$x'(n) = \frac{x(n)}{\max(|x|)}, \qquad 0 \le n \le N-1 \qquad (1)$$

2.3. Pre-emphasis

Proses pre-emphasis adalah proses yang dirancang untuk mengurangi dampak buruk dari transmisi dan suara latar. Proses pre-emphasis sangat baik dalam mengurangi efek distorsi, atenuasi, dan saturasi dari media rekaman. Perhitungan pre-emphasis dilakukan pada sinyal sinyal digital dalam domain waktu dan menggunakan persamaan (2).

$$\tilde{x}(n) = x'(n) - \alpha.x'(n-1) \quad (2)$$

2. 4. Framing dan Windowing



Gambar 3 Ilustrasi Proses *Framing* (Pengkerangkaan)

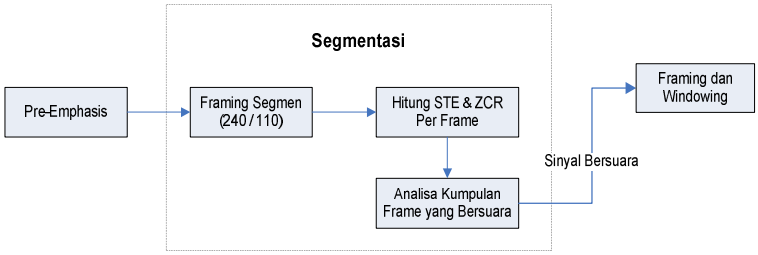
Framing atau pengkerangkaan adalah proses memecah sinyal digital ke dalam kelompok-kelompok waktu tertentu yang lebih singkat. Proses framing diawali dengan menentukan besar wilayah waktu setiap frame (Nw) dan besar wilayah pergeseran frame (Ns). Ilustrasi framing dapat dilihat pada Gambar 3. Sedangkan windowing adalah suatu perumusan untuk melemahkan nilai-nilai pada kedua ujung sinyal digital. Perumusan windowing yang digunakan adalah Hamming Windowing ditunjukkan pada persamaan (3) dan (4).

$$s(n) = x(n) * w(n) \quad (3)$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \quad (4)$$

2.5. Segmentasi

Pada penelitian ini segmentasi adalah proses memisahkan bagian bersuara dan bagian tidak bersuara. Hasil dari proses segmentasi adalah bagian sinyal *digital* bersuara. Pada penelitian ini, proses segmentasi dilakukan seperti bagan alir proses segmentasi pada Gambar 4.



Gambar 4 Bagan Alir Proses Segmentasi

Skenario kerja proses segmentasi adalah memecah sinyal digital ke dalam bentuk kerangka-kerangka atau frame-frame. Setiap frame memiliki 240 sampel sinyal dan pergeseran sampel sinyal untuk perpindahan frame sebesar 110 sampel sinyal. Pada langkah selanjutnya dilakukan pencarian nilai energy (STE) dalam persamaan (5) dan nilai zero crossing rate (ZCR) dalam persamaan (6) pada setiap frame. Suatu frame disebut sebagai bagian yang bersuara jika memenuhi suatu kondisi tertentu. Nilai kondisi yang digunakan pada penelitian ini diperoleh dari proses trial dan error. Kondisi yang dimaksud adalah sebagai berikut:

- Nilai STE pada frame lebih besar sama dengan 0.04
- Nilai ZCR pada frame lebih kecil sama dengan 80
- Memiliki minimal 4 frame secara berturut-turut yang memenuhi kedua kondisi di atas.

$$E = \frac{1}{N}\sum_{n=0}^{N-1}\left|x(n)\right| \tag{5}$$

$$Z = \frac{1}{2N}\sum_{n=1}^{N-1}\left|\mathrm{sgn}\{x(n)\} - \mathrm{sgn}\{x(n-1)\}\right| \tag{6}$$

$$\mathrm{sgn}\{x(n)\} = \begin{cases} 1 & x(n) \geq 0 \\ -1 & otherwise \end{cases} \tag{7}$$

2.6. Fast Fourier Transform (FFT)

Fast fourier transform merupakan pengembangan dari algoritma *discrete fourier transform* (DFT), yakni algoritma yang digunakan untuk mengubah sinyal *digital* pada *domain* waktu ke *domain* frekuensi. Proses FFT diperlukan untuk persiapan ekstraksi ciri MFCC. Perumusan FFT adalah:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}kn}, \;\; k = 0,1,...,N-1 \tag{8}$$

2.7. Ekstraksi Ciri LPC

Setelah sinyal mengalami proses framing dan windowing, langkah berikutnya adalah melakukan ekstraksi ciri LPC. LPC merupakan kependekan dari 'Linear Prediction Coding'. LPC merupakan salah teknik ekstraksi ciri yang sering digunakan dalam mengektraksi ciri sinyal digital suara. Akurasi kemampuan pengenalan pada penelitian sebelumnya, [3] dan [14], yang menggunakan ciri menunjukkan hasil yang baik sehingga mendorong penulis pada penelitian ini menggunakan ciri ini. Ada dua tahapan proses utama dalama melakukan ekstraksi ciri LPC, yakni proses autokorelasi dan proses analisis koefisien LPC.

Sinyal digital yang telah melalui proses *windowing* merupakan masukan dari proses autokorelasi. Pada proses autokorelasi perlu ditentukan suatu nilai orde analisis P (juga melambangkan banyak nilai ciri yang diambil). Nilai orde analisis ini biasanya memiliki nilai di antara 8 sampai 16. Pada penelitian ini, digunakan orde analisis 16. Perumusan untuk proses autokorelasi dapat dilihat pada persamaan (9).

$$r(p) = \sum_{n=1}^{N-p} s(n) * s(n+p) \tag{9}$$

Langkah proses selanjutnya setelah proses autokorelasi adalah melakukan analisis koefisien LPC. Nilai autokorelasi kemudian disusun dalam suatu matriks toeplitz seperti tampak pada persamaan (10). Penyelesaian dari matriks toeplitz dilakukan dengan menggunakan algoritma durbin seperti yang diuraikan pada persamaan (11) sampai persamaan (15).

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(P-1) \\ r(1) & r(0) & \cdots & r(P-2) \\ r(2) & r(1) & \cdots & r(P-3) \\ \vdots & \vdots & \ddots & \vdots \\ r(P-1) & r(P-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(P) \end{bmatrix} \tag{10}$$

$$E^{(0)} = r(0) \tag{11}$$

$$k_i = -\frac{\left\{r(i) + \sum_{j=1}^{i-1} a_j^{i-1} r(|i-j|)\right\}}{E^{(i-1)}}, \qquad 1 \leq i \leq P \tag{12}$$

$$a_j^{(i)} = k_i \tag{13}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i * a_{i-j}^{(i-1)} \tag{14}$$

$$E^{(i)} = \left(i - k_i^2\right)E^{(i-1)} \quad (15)$$

2.8. Ekstraksi Ciri MFCC

Umumnya persepsi frekuensi pendengaran manusia pada dasarnya tidaklah linear seperti pada pemodelan frekuensi umumnya. Pemodelan persepsi ini kemudian diberinama mel-scale dan frekuensinya disebut frekuensi mel. Perumusan untuk mengkonversi frekeunsi biasa (linear) ke frekuensi mel dan sebaliknya ditunjukkan persamaan (16) dan (17).

$$\hat{f}_{mel} = 1127 * \ln\left(1 + \frac{f_{lin}}{700}\right) \quad (16)$$

$$f_{lin} = \hat{f}_{mel}^{-1} = 700 * \left[\exp\left(\frac{\hat{f}_{mel}}{1127}\right) - 1\right] \quad (17)$$

Proses ekstraksi ciri MFCC di mulai dengan menentukan jumlah filter yang akan digunakan. Jumlah filter yang digunakan biasanya berkisar di antara nilai 20 hingga 40. Dari jumlah filter ini kemudian akan dibangkitkan sejumlah filterbank sesuai dengan persamaan (18) dan (19).

$$f_b(m) = \hat{f}_{mel}^{-1}\left(\hat{f}_{mel}\left(f_{lin-low}\right) + m.\frac{\hat{f}_{mel}\left(f_{lin-high}\right) - \hat{f}_{mel}\left(f_{lin-low}\right)}{P+1}\right) \quad (18)$$

$$H_m(k) = \begin{cases} 0 & for \quad k < f_b(m-1) \\ \dfrac{k - f_b(m-1)}{f_b(m) - f_b(m-1)} & for \quad f_b(m-1) \le k \le f_b(m) \\ \dfrac{f_b(m+1) - k}{f_b(m+1) - f_b(m)} & for \quad f_b(m) \le k \le f_b(m+1) \\ 0 & for \quad k > f_b(m+1) \end{cases} \quad (19)$$

Setelah filterbank terbentuk, langkah berikutnya adalah melakukan mel frequency warping untuk memperoleh nilai log-mel frequency cepstrum. Perumusan mel frequency warping dapat dilihat pada persamaan (20).

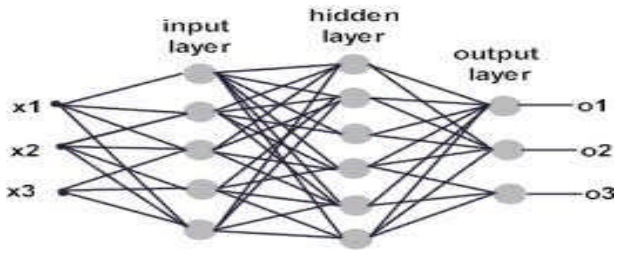
$$a\_mfc_n^m(k) = \ln\left(\sum_{k=1}^{K} |X_n(k)|^2 H_m(k)\right) \quad (20)$$

Langkah terakhir ekstraksi ciri dari proses ekstraksi ciri adalah menggunakan perumusan discrete cosine transform untuk memperoleh nilai koefisien MFCC-nya. Pada tahap terakhir ini kita dapat menentukan jumlah koefisien yang akan digunakan sebagai ciri. Jumlah koefisien Q minimum yang umumnya diambil untuk proses pengenalan suara adalah 13 dan maksimum sejumlah filter yang didefinisikan. Perumusan untuk persolehan nilai koefisien MFCC ditunjukkan pada persamaan (21).

$$c_n^q = \sum_{m=1}^{P} a\_mfc_n^m . \cos\left[m\left(q - \frac{1}{2}\right)\frac{\pi}{P}\right] \quad (21)$$

2.9. Jaringan Syaraf Tiruan Back-Propagation

Jaringan syaraf tiruan *back-propagation* adalah suatu metode untuk mengklasifikasi dan mengidentifikasi suatu pola masukan tertentu dengan melakukan perbaikan bobot penghantar antar lapisan. Secara definitif, Fausett [12] mendefinisikan jaringan syaraf tiruan sebagai sistem pemrosesan informasi yang mempunyai karakteristik kinerja tertentu seperti pada jaringan syaraf biologis manusia. Secara sederhana, bentuk dari jaringan syaraf tiruan *back-propagation* tampak pada Gambar 5.



Gambar 5 Ilustrasi Arsitektur JST

Pada penelitian ini, peneliti menggunakan bantuan *tools* dari *software* MATLAB 6.5 untuk memodelkan jaringan syaraf tiruan *backpropagation*. Pembahasan tetntang *tools neural network* yang terdapat pada *software* MATLAB 6.5 dapat dilihat pada buku Demuth dan Beale [15]. Adapun beberapa parameter yang mendapat pengaturan tampak pada Tabel 1.

Tabel 1 Nilai Parameter JST *Backpropagation*

No	Parameter yang diatur	Nilai Parameter	Keterangan
1	Jumlah Neuron Input	T2 ← 12	LPC *Frm* Tunggal
		T2 ← 588	LPC *Frm* Banyak
		T1 ← 20	MFC*Frm* Tunggal
		T1 ← 980	MFC*Frm* Banyak
2	Jum Hidden Layer	2	
3	Jum Neuron Hidden	[400 200]	
4	Jum Neuron Output	11	
5	Perbaikan Bobot	Trainrp	*Resillent BP*
6	Fungsi Aktivasi	Logsig	Logsig
7	Error Minimum	0.0001	
8	Maksimal Epoch	10000	

Hal berikutnya yang diatur penulis adalah target keluaran. Target keluaran jaringan syaraf tiruan yang terdiri dari 1741 suku kata bahasa lisan Bahasa Indonesia diatur dalam bentuk biner.

3. HASIL DAN PEMBAHASAN

Dalam rangka pengujian untuk mengetahui kemampuan pengenalan 1741 suku kata bahasa lisan Bahasa Indonesia, penulis melakukan pengujian menggunakan rancangan sistem seperti yang tampak pada Gambar 1 dan 2. Beberapa bagian pada rancangan sistem memerlukan nilai parameter yang pasti sehingga penulis menentukan nilai parameter tersebut. Nilai parameter yang dimaksud adalah nilai koefisien pre-emphasis: 0.97, jumlah orde analisis koefisien LPC: 12, jumlah filterbank pada ekstraksi ciri MFCC: 20, dan jumlah koefisien cepstral MFCC: 20. Nilai parameter yang ditentukan merupakan hasil penelitian awal yang menunjukkan nilai pengenalan yang lebih baik. Pengujian pada sistem ini menggunakan bantuan hardware berupa satu unit laptop Axioo Neon MNC T6400 standar yang dilengkapi dengan alat perekam suara (mic Sennheisser PC 110), software MATLAB 6.5, dan software CoolEdit Pro 2.1. Pada penelitian ini, penulis lebih menekankan eksplorasi pada parameter *framing*, jumlah target data, kemiripan bunyi ucapan suku kata, dan jumlah data pelatihan untuk

menjelaskan hal-hal yang mempengaruhi kemampuan pengenalan terhadap 1741 suku kata bahasa lisan Bahasa Indonesia. Setiap pengujian dilakukan terhadap data latih sebanyak tiga kali dan diambil hasil pengenalan terbaiknya. Setelah itu pengujian dilakukan terhadap data uji.

Observasi Utama

Pengujian utama ini adalah pengenalan terhadap 1741 suku kata bahasa lisan Bahasa Indonesia. Hasil pengujian disajikan pada Tabel 3.

Tabel 3 Hasil Pengujian 1741 Suku Kata Bahasa Lisan

Data Ciri	Kemampuan Pengenalan (%)	
	Data Latih	Data Uji
MFCC_1	85.75	0.65
MFCC_2	59.46	0.33
LPC_1	95.80	0.42
LPC_2	70.04	0.52

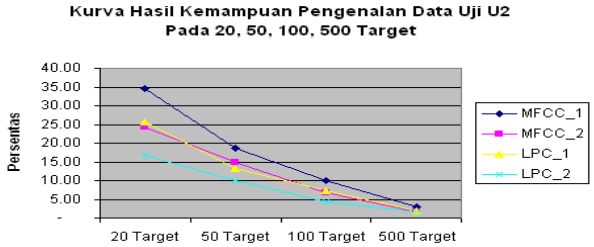
Pada Tabel 3 terlihat data ciri menggunakan kode 1 dan 2 di akhir dari nama ciri. Pengkodean angka 1 berarti ciri tersebut menggunakan framing dengan ukuran 30ms dan pergeseran 20ms. Sedangkan kode 2 berarti ciri tersebut menggunakan framing tunggal. Asumsi waktu maksimal ucapan lafal kata adalah 1 detik, maka akan diperoleh 49 frame setiap ucapan lafal kata pada framming 30/20. Asumsi ini menghasilkan 980 nilai ciri MFCC dan 588 nilai ciri LPC pada ciri dengan kode 1, sedangkan pada ciri dengan kode 2 menghasilkan 20 nilai ciri MFCC dan 12 nilai ciri LPC (sesuai dengan neuron input JST Tabel 2).

Dari hasil pengujian pada Tabel 3, hasil pengenalan terhadap data latih menunjukkan hasil yang sangat baik, dengan mencapai angka 95.80% pada ciri LPC dengan frame banyak (LPC_1). Hal ini menunjukkan bahwa penggunaan identifier jaringan syaraf tiruan tepat. Pada tahap generalisasi ini, jaringan syaraf tiruan dapat menemukan kesamaan pola masukan data pelatihan dengan sangat baik. Akan tetapi keberhasilan jaringan syaraf tiruan untuk mengeneralisasi, tidak diikuti dengan kemampuan pengenalan yang baik pada pengujian pengenalan yang seharusnya (data uji). Kemampuan pengenalan terbaik pada tahap pengujian data uji hanya mencapai angka 0.65%. Angka ini berbanding terbalik dengan keberhasilan pengenalan kembali. Mengapa hal ini bisa terjadi? Padahal beberapa parameter telah diatur pada nilai yang memberikan kontribusi pengenalan yang paling baik. Padahal hasil pengenalan kembali sangat tinggi. Apakah yang menyebabkan hal ini terjadi? Pada pengamatan lainnya, eksploarasi *framing* 30/20 dan tunggal belum menunjukkan hasil yang konstan sehingga belum dapat disebutkan pengggunaan *framing* mana yang lebih baik.

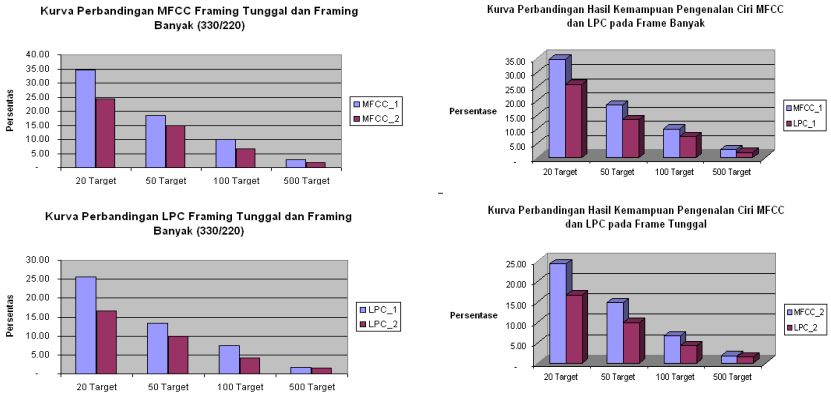
Observasi Bantuan 1 (Observasi pengaruh jumlah target)

Pertanyaan mengapa hasil pengenalan pada tahap pengujian 1741 target tidak baik coba dijawab penulis dengan mengeksplorasi jumlah target data. Jumlah target data yang berjumlah 1741 dipecah menjadi 20, 50, 100, dan 500 target. Pemilihan target dilakukan secara random. Adapun hasil pengujian disajikan pada Gambar 7.

Bentuk kurva hasil pengujian dengan jumlah target 20, 50, 100, dan 500 pada Gambar 7 menunjukkan adanya penurunan akurasi pengenalan seiring dengan bertambahnya jumlah target. Hal ini memberikan penjelasan awal mengapa hasil pengujian terhadap 1741 suku kata bahasa lisan Bahasa Indonesia menunjukkan hasil yang tidak memuaskan. Hasil pengamatan yang memberikan pertanyaan baru kepada penulis adalah akurasi kemampuan pengenalan terbaik yang hanya mencapai angka 35% dengan jumlah target 20. Akurasi pengenalan yang berada pada angka 35% tidaklah sesuai dengan akurasi pengenalan pada penelitian lain yang menggunakan ciri serupa pada [3] dan [14]. Hal ini akan diselidiki lebih jauh pada penelitian ini dengan mengeksplorasi kemiripan bunyi pengucapan suku kata bahasa lisan dan jumlah target yang dilatihkan.



Gambar 7 Kurva Hasil Pengujian dengan Jumlah Target 20, 50, 100, dan 500

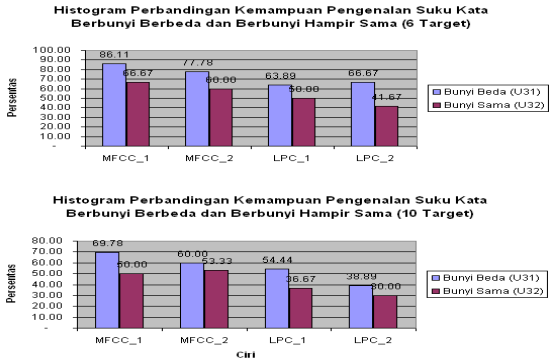


Gambar 8 Histogram Perbandingan Kemampuan Pengenalan Antara Framing 30/20-Framing Tunggal (kiri) dan Antara Ciri MFCC-Ciri LPC (kanan)

Pengamatan lainnya berkaitan lainnya berkaitan dengan hasil observasi ini adalah pada perbandingan kemampuan pengenalan dengan menggunakan teknik *framing* 30/20 dengan *framing* tunggal dan perbandingan kemampuan pengenalan antar ciri MFCC, LPC. Kedua perbandingan tersebut disajikan pada Gambar 8. Pada Gambar 8 sebelah kiri ditunjukkan kemampuan pengenalan dengan menggunakan *framing* 30/20 (biru) lebih baik dibandingkan dengan *framing* tunggal (merah). Nilai ciri pada *framing* 30/20 lebih sensitif terhadap perubahan sinyal pada waktu tertentu. Hal ini dapat menjelaskan bahwa nilai ciri dengan memenggal sinyal pada satuan waktu tertentu dan mengumpulkannya lebih baik dibandingkan dengan menggunakan ciri pada keseluruhan sinyal. Sedangkan hasil pada Gambar 8 sebelah kanan menunjukkan bahwa kemampuan pengenalan dengan menggunakan ciri MFCC (biru) lebih baik dari ciri LPC (merah). Kedua ciri ini merepresentasikan ciri pada domain yang berbeda, yakni domain frekuensi (MFCC) dan domain waktu (LPC). Dari hasil ini, setidaknya dperoleh kesimpulan bahwa penggunaan ciri MFCC lebih baik dari ciri LPC pada kasus ini.

Observasi Bantuan 2 (Observasi pengaruh bunyi)

Observasi untuk menjawab perbedaan akurasi pengenalan yang jauh oleh ciri MFCC, LPC pada penelitian ini dengan penelitian lain coba dilakukan dengan menggunakan jumlah target yang lebih kecil lagi (seperti pada penelitian lain yang mengklaim akurasi di atas 88%). Jumlah target yang digunakan adalah 6 dan 10 target. Data yang digunakan masih merupakan data utama. Jumlah target 6 dan 10 ini dibedakan lagi menurut bunyi ucapan suku katanya, yakni yang berbunyi beda dan yang berbunyi hampir sama. Perbedaan ini diyakini oleh penulis menjadi salah satu penyebab tidak baiknya akurasi pengenalan seperti yang tampak pada hasil pengujian Tabel 1 dan 2. Contoh 6 target dengan bunyi ucapan berbeda adalah /a/, /i/, /u/, /ə/, /e/, /o/ dan penambahan bunyi /ai/, /au/, /ei/, dan /oi/ dari 6 target berbunyi berbeda tersebut untuk 10 target. Sedangkan contoh target dengan bunyi hampir sama adalah /a/, /an/, /am/, /ah/, /at/, /ak/, /ap/, /as/, /al/, dan /ar/.

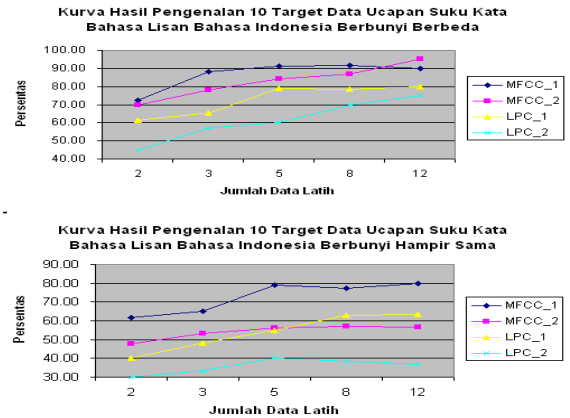


Gambar 9 Histogram Perbandingan Kemampuan Pengenalan Antara Data Ucapan Berbunyi Beda dan Berbunyi Hampir Sama

Hasil pengamatan terhadap pengujian eksplorasi data ucapan yang berbunyi berbeda dan hampir sama menunjukkan adanya pengaruh data ucapan yang berbunyi berbeda dan berbunyi hampir sama (lihat Gambar 9). Kemampuan pengenalan terhadap data ucapan dengan bunyi berbeda (biru) menunjukkan hasil yang lebih baik dibandingkan kemampuan pengenalan terhadap data ucapan dengan bunyi hampir sama (merah). Fakta ini didukung pula bahwa manusia dalam mempersepsikan ucapan yang memiliki bunyi berbeda memang lebih baik dibanding dengan ucapan yang berbunyi hampir sama. Akurasi kemampuan pengenalan juga telah mencapai angka yang baik yakni 86% untuk 6 target (MFCC_1) dan 70% untuk 10 target (MFCC_1). Hasil pengenalan ini menunjukkan bahwa kemampuan ciri yang digunakan telah sesuai dengan hasil penelitian terdahulu yang menggunakan ciri yang sama. Hasil pengujian pada bagian ini juga memberikan fakta penjelasan tambahan terhadap kegagalan pengenalan terhadap data utama (1741 target), yakni data utama yang terlalu banyak mengandung kesamaan bunyi yang harus dikenali sehingga akurasi pengenalan menjadi turun.

Observasi Bantuan 3 (Observasi pengaruh jumlah data latih)

Observasi berikutnya adalah mengetahui pengaruh pertambahan jumlah data pelatihan terhadap kemampuan pengenalan. Observasi ini dilakukan penulis, dengan harapan memberikan gambaran yang lebih jelas terhadap kekurangan dari sistem sekaligus memberikan pertimbangan tambahan dalam penelitian ke depannya berkaitan dengan jumlah data latih yang digunakan. Eksplorasi dilakukan terhadap jumlah target yang lebih kecil yakni 10 target. Data yang digunakan pada observasi pengaruh pertambahan jumlah data pelatihan ini juga berbeda, yakni rekaman pengucapan 10 target suku kata bahasa lisan Bahasa Indonesia berbunyi berbeda dan hampir sama hanya berasal dari satu orang saja. Total rekaman akan berisi 15 kali ucapan 10 target suku kata berbunyi berbeda dan 15 kali ucapan 10 target suku kata berbunyi hampir sama. Dari rekaman tersebut dibagi ke dalam beberapa kelompok pengujian yakni pelatihan dengan banyak data latih 2, 3, 5, 8, dan 12 rekaman diikuti pengujian sisa data rekaman sebagai data uji.



Gambar 10 Kurva Hasil Pengenalan dengan Menggunakan Jumlah Data Latih 2, 3, 5, 8, dan 12

Dari hasil pengujian identifikasi 10 target pada beberapa kelompok jumlah data yang dilatihkan menunjukkan adanya trend peningkatan kemampuan pengenalan seiring dengan bertambahnya jumlah data yang dilatihkan (kurva pada Gambar 10). Angka akurasi pengenalan juga mencapai angka yang sangat baik yakni 90% (MFCC_1). Hasil observasi ini menunjukkan bahwa penambahan data latih akan meningkatkan akurasi pengenalan. Hal ini dikarenakan semakin banyak data latih berarti semakin banyak pola yang dapat dipelajari. Semakin banyak pola yang dipelajari berarti pengetahuan tentang suara semakin baik. Pada akhirnya akurasi pengenalan semakin tinggi. Walaupun demikian peningkatan jumlah data latih berarti semakin meningkatnya kebutuhan memori dan komputasi untuk pelatihan yang akan memperlambat pelatihan dan bahkan terbentur batas atas memori yang digunakan. Oleh karena itu, jika penambahan jumlah data latih menjadi solusi, maka perlu pertimbangan matang untuk permasalahan hardware.

4. KESIMPULAN

Berdasarkan pada hasil penelitian, pendekatan teknik pengolahan sinyal yang dilakukan penulis untuk mengenali suku kata bahasa lisan Bahasa Indonesia ternyata tidak mampu memberikan akurasi pengenalan yang baik (0.65%). Hal ini ditunjukkan pada observasi utama penelitian ini. Pendekatan ekstraksi ciri MFCC dan LPC tidak cocok untuk mengenali suara dengan jumlah target yang besar (1741 target), namun ekstraksi ciri ini cocok untuk mengenali suara dengan jumlah data yang kecil. Selain jumlah target yang besar, juga ada permasalahan terhadap kesamaan bunyi yang dimiliki pada jumlah target yang besar (observasi bantuan 2).

Adapun beberapa fakta lain yang ikut muncul sebagai hasil penelitian pengenalan suara pengucapan suku kata bahasa lisan Bahasa Indonesia adalah sebagai berikut.

1. Kemampuan pengenalan sangat dipengaruhi oleh jumlah target yang digunakan. Semakin besar jumlah target yang harus dikenali maka kemungkinan penurunan kemampuan pengenalan semakin besar pula Pernyataan ini dapat dilihat pada observasi bantuan 1 dan 2.
2. Kemampuan pengenalan dengan menggunakan kerangka banyak lebih baik dibandingkan dengan kerangka tunggal karena kerangka banyak lebih sensitif terhadap perubahan sinyal pada wilayah waktu tertentu (observasi bantuan 1, 2, dan 3).
3. Kemampuan pengenalan dengan menggunakan ciri MFCC relatif lebih baik dibandingkan menggunakan ciri LPC (observasi bantuan 1, 2, dan 3).
4. Kemampuan pengenalan cenderung semakin baik ketika jumlah data pelatihan bertambah (observasi bantuan 3).

5. SARAN

Dalam rangka memperbaiki kemampuan pengenalan terhadap suku kata bahasa lisan Bahasa Indonesia, penulis menyarankan untuk melakukan klasifikasi awal terlebih dahulu pada penelitian lebih lanjut. Klasifikasi awal dapat dimulai dengan mensurvei perbedaan bunyi suku kata. Hal ini didasarkan pada hasil pengujian yang menunjukkan kemampun pengenalan yang lebih baik pada jumlah target yang kecil dan suku kata dengan bunyi berbeda. Selain dengan menggunakan klasifikasi awal, cara lain yang dapat dilakukan adalah menguji kemampuan ciri lain sebagai wakil dari sinyal digital suara ucapan suku kata bahasa lisan Bahasa Indonesia.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Sekolah Tinggi Manajemen Informatika dan Komputer Widya Dharma yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] Sitanggang, D., Sumardi, Hidayatno, A., 2002, Pengenalan Vokal Bahasa Indonesia dengan Jaringan Syaraf Tiruan Melalui Transformasi Fourier, *Seminar Nasional I Rekayasa Aplikasi dan Industri (RAPI)*, Semarang.

[2] May, I.L., Sumardi, Hidayatno, A., 2002, Pengenalan Vokal Bahasa Indonesia dengan Jaringan Syaraf Tiruan Melalui Transformasi Wavelet, *Seminar Nasional I Rekayasa Aplikasi dan Industri (RAPI)*, Semarang.

[3] Amri, Arhami M., Fauzan, 2008, Analisa Teknik Pengenalan Sinyal Wicara dengan Hidden Markov Models - Neural Network, *Jurnal Listrik Telekomunikasi dan Elektronika 2008 (LiTEk)*, No.2, Vol. 5, Hal 64-67.

[4] Ajulian, A.Z., Hidayatno, A., Widyanto, M.T.S., 2008, Aplikasi Pengenalan Ucapan Sebagai Pengatur Mobil Dengan Pengendali Jarak Jauh, *Jurnal Berkala Transmisi Teknik Elektro*, No.1, Jilid 10, Hal 21-26.

[5] Hapsari, J.P., 2007, Aplikasi Pengenalan Suara dalam Pengaksesan Sistem Informasi Akademik, *Skripsi S-1*, Fakultas Teknik Elektro, Universitas Diponegoro, Semarang.

[6] Rizal, A., Suryani, V., 2008, Pengenalan Signal EKG Menggunakan Dekomposisi Paket Wavelet dan K-Means Clustering, *Seminar Nasional Aplikaso Teknologi Informasi 2008 (SNATI)*, Yogyakarta, 21 Juni 2008

[7] Rabiner L., Juang B.H., 1993, *Fundamentals for Speech Recognition*, Prentice Hall, New Jersey.

[8] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D., 2008, Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy, Khaled Elleithy, *Advanced Techniquee in Computing Sciences and Software Engineering*, *Springer Science+Business Media*, New York. Hal. 279-282.

[9] Oppenheim, A.V., Schafer, R.W., 1999, *Discrete Time Signal Processing*, Second Edition, Prentice Hall, New Jersey.

[10] Gonzales R.C., Woods R.E., 2008, *Digital Image Processing, 3rd Edition*, Prentice Hall, New Jersey.

[11] Huang, X.D., Acero, A., Hon, H.W., 2001, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, New Jersey.

[12] Fausett L., 1994, *Fundamentals of Neural Networks: Architecture, Algorithms, dan Applications*, Prentice Hall, New Jersey.

[13] Pusat Pengembangan Bahasa Indonesia, 2000, *Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan*, Pusat Bahasa Departemen Pendidikan Nasional, Jakarta.

[14] Ahmad, A.M., Ismail, S., Samaon, D.F., 2004, Reccurent Neural Network with Backpropagation through Time for Speech Recognition, *International Symposium on Communications and Information Technologies 2004 (ISCIT)*, Sapporo (Jepang), 26-29 Oktober 2004.

[15] Demuth H, Beale M., 2000, *Neural Network Toolbox: For Use with MATLAB*, The MathWorks, Natick Massachusetts USA.