

Pengenalan Suara Vokal Bahasa Indonesia dengan Jaringan Saraf Tiruan Menggunakan Ciri Transformasi Wavelet Diskrit

Nadya Amalia¹, Arfan E Fahrudin² dan Amar V Nasrullah²

Abstract: Vowel recognition is the main topic in speech recognition. There are six Indonesian vowels, i.e. /a/, /i/, /u/, /e/, /ə/ and /o/. Feature extraction is an important step in recognition system because the recognition rate depends on feature extraction results. Vowel feature extraction via discrete wavelet transform (DWT) is presented here. Mother wavelet db4 and sym4 are used. Minimum, maximum, mean and standard deviation value of wavelet coefficients are extracted as vowel features. DWT with level 2 decomposition obtains 12 features, level 4 decomposition obtains 20 features and level 6 decomposition obtains 28 features. Then, those vowel features are used as an input of artificial neural network (ANN) with 2 hidden layers. First hidden layer has 10 neurons and second hidden layer has variety 5 and 7 neurons. Backpropagation method is used to train the ANN. The vowel signals are recorded from 10 female respondents and 10 male respondents. Each respondent pronounces six Indonesian vowels and syllable /ka/, /ki/, /ku/, /ke/, /kə/ and /ko/. Experimental results show that the best recognition rate for the vowel is 85%, which is obtained by using mother wavelet sym4, level 6 decomposition and 7 neurons for second hidden layer, and the best recognition rate for the syllable is 80%, which is obtained by using mother wavelet db4, level 6 decomposition and 5 neurons for second hidden layer.

Keywords: Indonesian vowels, discrete wavelet transform, artificial neural network, backpropagation

PENDAHULUAN

Suara merupakan alat komunikasi paling mendasar bagi manusia. Saat ini, suara tidak hanya digunakan untuk komunikasi antar manusia tetapi juga digunakan untuk berkomunikasi dengan perangkat-perangkat teknologi, sehingga topik pengenalan suara pun semakin berkembang. Pengenalan suara vokal adalah sub topik yang mendasar dalam bidang pengenalan suara.

Secara umum ada enam suara vokal dalam bahasa Indonesia, yaitu: /a/, /i/, /u/, /e/, /ə/ dan /o/ [1].

Ciri yang dimiliki suara biasanya dapat diperoleh melalui representasi waktu-frekuensi. Proses untuk mendapatkan ciri tersebut dikenal sebagai ekstraksi ciri. Ekstraksi ciri merupakan salah satu langkah penting dalam sistem pengenalan karena tingkat akurasi pengenalan tergantung pada ciri yang diekstraksi [2]. Pada

¹Mahasiswa dan ²Staf Pengajar Program Studi Fisika, Universitas Lambung Mangkurat

penelitian ini, ekstraksi ciri suara vokal bahasa Indonesia dilakukan dengan menggunakan transformasi wavelet diskrit (TWD). Kelebihan yang dimiliki TWD adalah kemampuannya untuk melakukan analisis multiresolusi, yakni melakukan analisis waktu dan frekuensi secara bersamaan yang sesuai digunakan untuk menganalisis suara yang pada dasarnya merupakan sinyal *non-stasioner* [3].

Pengenalan dilakukan dengan jaringan saraf tiruan (JST). Seperti pada otak manusia, JST terdiri atas *neuron-neuron* yang saling berhubungan yang dapat bekerja sama satu dengan yang lainnya untuk membentuk suatu sistem. Fitur utama yang menjadi alasan digunakannya JST adalah kemampuan untuk belajar dari data pelatihan dan generalisasi ke situasi/kondisi yang baru [4]. Pada penelitian ini, nilai minimum, nilai maksimum, *mean* dan standar deviasi dari koefisien-koefisien *wavelet* diekstraksi sebagai ciri suara vokal yang kemudian digunakan sebagai *input* untuk JST. JST dilatih dengan menggunakan metode pembelajaran *backpropagation*.

Suara Vokal Bahasa Indonesia

Suara vokal atau yang juga dikenal sebagai *voiced sound* adalah

bunyi bahasa yang arus udaranya tidak mengalami rintangan. Secara umum ada enam suara vokal dalam bahasa Indonesia, yaitu: /a/, /i/, /u/, /e/, /ə/ dan /o/ [1]. Kualitas suara vokal ditentukan oleh tiga faktor seperti yang ditunjukkan pada **Tabel 1**.

Tabel 1. Vokal bahasa Indonesia [5]

Bentuk Bibir	Tak Bundar		Bundar
	Depan	Tengah	Belakang
Bagian Lidah Tinggi	i		u
Sedang	e	ə	o
Rendah	A		

Transformasi Wavelet Diskrit

Transformasi wavelet Mallat [6] memiliki kemampuan untuk melakukan analisis multiresolusi. Menurut Addison [7] transformasi wavelet didefinisikan sebagai:

$$X_{WT}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t) \psi * \left(\frac{t-\tau}{s} \right) dt \dots (1)$$

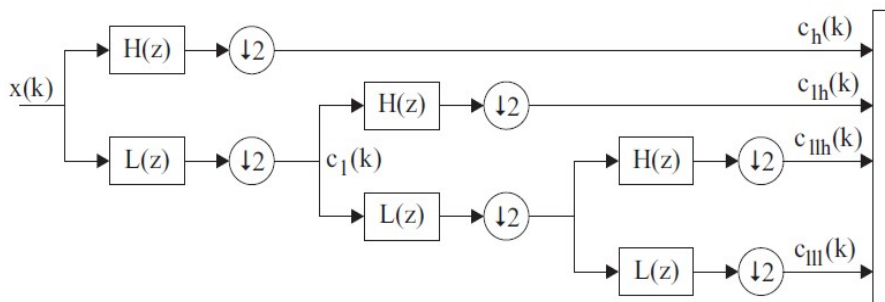
Sinyal $X_{WT}(\tau, s)$ merupakan fungsi dari parameter translasi τ dan parameter skala s . ψ menunjukkan *mother wavelet*, yakni gelombang kecil yang digunakan untuk melakukan analisis [7].

Transformasi wavelet diskrit (TWD) bukanlah transformasi wavelet yang didiskritisasi. TWD melakukan analisis multiresolusi dengan menggunakan *filter bank* [8]. *Filter bank* terdiri atas *filter-filter* yang membagi sinyal menjadi beberapa *band* frekuensi

yang juga sering disebut sebagai proses dekomposisi [9].

Dekomposisi dapat dikembangkan menjadi beberapa level, tergantung pada resolusi yang diinginkan [8]. Secara teori tidak ada batasan dalam penentuan level tersebut. Namun dalam kenyataannya,

proses tersebut hanya dapat berlangsung hingga masing-masing koefisien detail mengandung sebuah *sample* dari sinyal terdekomposisi saja. Pada prakteknya, penentuan level dapat didasarkan pada kebutuhan dan sifat-sifat alami sinyal [10].



Gambar 1. Dekomposisi level tiga [8]

Sinyal-sinyal *output* yang berbeda dari *analysis filter bank* disebut *subband*, dan teknik *filter bank* juga dikenal sebagai *subband coding*. *Subband c_{III}* disebut juga koefisien aproksimasi (cA) dan *subband-subband* yang lain disebut koefisien detail (cD). Untuk dekomposisi level-*p*, konten frekuensi dari cA dan cD dapat dihitung sebagai [11]:

$$f_{cA} = [0, 2^{-p-1}f_s] \dots\dots(2)$$

$$f_{cDp} = [2^{-p-1}f_s, 2^{-p}f_s] \dots\dots(3)$$

Sementara itu, nilai cA dan cD dapat dihitung dengan:

$$cA[p] = \sum_n x[k]L[2p - k] \dots\dots(4)$$

$$cD[p] = \sum_n x[k]H[2p - k] \dots\dots(5)$$

Jaringan Saraf Tiruan

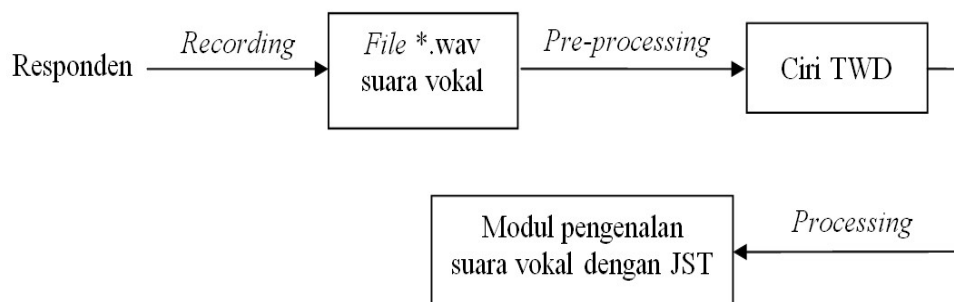
Jaringan saraf tiruan (JST) adalah suatu metode komputasi yang meniru sistem jaringan saraf biologis [12]. JST mempunyai kemampuan *memorisasi*, yaitu kemampuan untuk mengenali kembali sebuah pola yang telah dipelajari, dan kemampuan *generalisasi*, yaitu kemampuan untuk menghasilkan respon yang bisa diterima terhadap pola-pola *input* yang serupa (namun tidak identik) dengan pola-pola yang sebelumnya telah dipelajari. [13]

Backpropagation merupakan metode pembelajaran untuk melatih JST yang termasuk dalam kategori *supervised learning*. Sesuai dengan

namanya, nilai *error* pada *backpropagation* dipropagasi balik dari *node output* menuju *node* di sebelah dalam. Jadi, secara teknis dapat dikatakan bahwa *backpropagation* digunakan untuk menghitung *gradient* dari nilai-nilai *error* dalam jaringan dengan memperhatikan bobot jaringan yang dapat diubah-ubah [14].

METODE PENELITIAN

Penelitian pengenalan suara Vokal Bahasa Indonesia dengan Jaringan Saraf Tiruan dibagi menjadi tiga tahapan, yaitu perekaman sampel suara, *pre-processing* dan *processing*. Tahapan penelitian seperti diagram *block* yang ditunjukkan oleh **Gambar 2**.



Gambar 2. Diagram Blok Tahapan Penelitian

1. Perekaman Sampel Suara

Perekaman sampel suara dilakukan terhadap 20 responden yaitu 10 orang perempuan dan 10 orang laki-laki dengan usia 19-25 tahun. Setiap responden mengucapkan keenam suara vokal bahasa Indonesia (/a/, /i/, /u/, /ə/, /e/ dan /o/) dan suara vokal dengan tambahan konsonan /k/ (/ka/, /ki/, /ku/, /kə/, /ke/ dan /ko/). Perekaman dilakukan selama 3s dengan menggunakan *Audacity 2.0* melalui *microphone* dalam format *.wav* dengan frekuensi *sampling* $f_s = 44,100\text{Hz}$. Untuk setiap suara vokal dan suku kata dilakukan perekaman sebanyak dua

kali. Rekaman sampel suara dibagi menjadi dua, data dari 5 responden perempuan dan laki-laki pertama digunakan untuk pelatihan jaringan dan sisanya untuk pengujian jaringan.

2. Tahap *Pre-processing*

Gambar 3 merupakan diagram alir dari tahapan *pre-processing*, yaitu:

a. *Sampling*

Teori Shannon menyatakan bahwa frekuensi *sampling* minimum adalah 2 kali frekuensi sinyal analog [15]. Frekuensi efektif suara manusia adalah 4,000 Hz dan frekuensi *sampling* yang digunakan pada penelitian ini adalah 44,100 Hz.

b. Pre-emphasis filter

Pre-emphasis filter berguna untuk menghaluskan bentuk spektral dari sinyal suara. Filter ini mempunyai fungsi transfer [16]:

$$H(z) = 1 - az^{-1} \dots\dots\dots(6)$$

Dimana a merupakan konstanta pre-

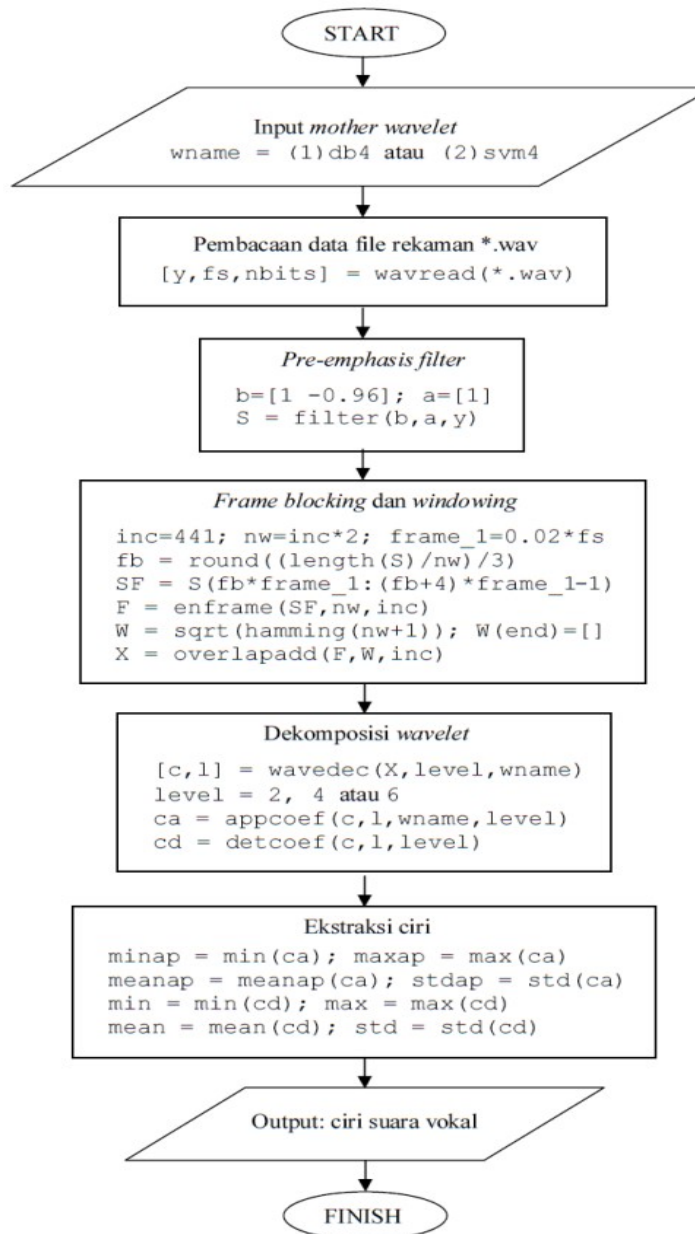
emphasis filter, biasanya bernilai $0.9 \leq a \leq 1$.

c. Frame Blocking dan Windowing

Sinyal yang telah di-filter kemudian dibagi-bagi menjadi frame-frame sepanjang 20 ms. Perhitungan ini ditunjukkan oleh persamaan 7 dan 8.

(Sampling Rate) (Frame Length = Number of Samples in a Frame) (7)

(44100 sampel/s)(0.02 s) = 882 sampel(8)



Gambar 3. Flowchart tahap pre-processing

Frame blocking bertujuan agar hasil rekaman sinyal suara yang merupakan *time variant* bisa dianggap menjadi *time invariant* sehingga proses komputasi dapat dilakukan dengan lebih cepat. Sementara itu, *windowing* dilakukan pada setiap *frame* dengan tujuan untuk meminimalisasi diskontinuitas sinyal pada awal dan akhir *frame*. Persamaan 9 merupakan persamaan yang membangkitkan *Hamming window*.

$$W(n) = 0.54 - 0.46 \cos(2\pi n / 881) \dots (9)$$

d. Dekomposisi TWD

Mother wavelet yang digunakan pada penelitian ini adalah *db4* dan *sym4* dengan level dekomposisi 2, 4 dan 6. Dekomposisi TWD menghasilkan koefisien-koefisien wavelet, di mana nilai minimum, maksimum, *mean* dan standar deviasi (4 parameter) dari koefisien-koefisien wavelet tersebut diekstraksi sebagai ciri suara vokal yang kemudian dijadikan sebagai *input* untuk JST.

3. Tahap Processing

Diagram alir tahapan **Processing** seperti yang ditunjukkan oleh **Gambar 4** dan **Gambar 5**.

a. Membangun Arsitek JST.

Arsitektur JST yang digunakan pada penelitian ini adalah *multilayer feed-forward backpropagation* dengan 2 *hidden layer* dan memvariasikan jumlah *neuron* pada *hidden layer*. Jumlah *neuron* pada *hidden layer* pertama adalah 10 dan pada *hidden layer* kedua divariasikan 5 dan 7. Fungsi aktivasi *sigmoid bipolar* digunakan pada kedua *hidden layer* dan fungsi aktivasi *linear* digunakan pada *output layer*.

b. Pelatihan Jaringan

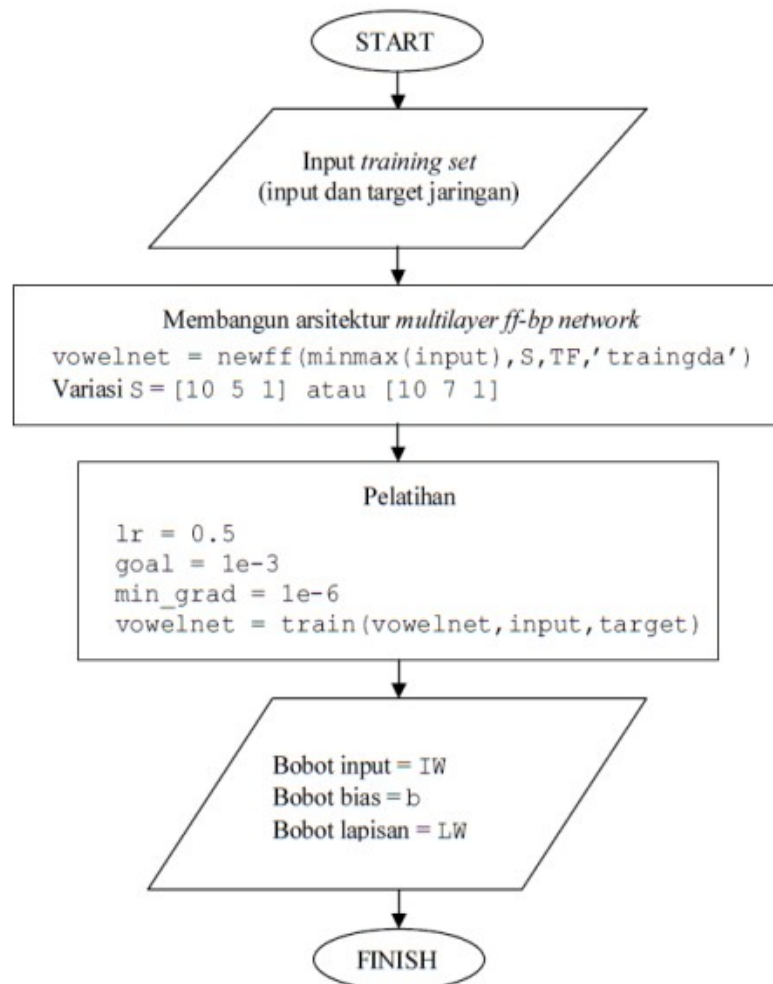
Pelatihan ini bertujuan untuk mendapatkan nilai bobot dan bias yang paling sesuai untuk jaringan sehingga *input* jaringan mampu menghasilkan *output* yang sesuai dengan target.

Berhubung tidak ada aturan baku dalam penentuan nilai target suatu jaringan, maka pada penelitian ini target jaringan ditentukan sendiri dengan catatan masing-masing suara vokal memiliki nilai target yang berbeda untuk membedakan target dari *input* ciri suara vokal yang satu dengan target dari *input* ciri suara vokal yang lain. Untuk suara vokal /a/, /i/, /u/, /ə/, /e/ dan /o/ targetnya berturut-turut adalah 1, 3, 5, 7, 9 dan 11. Dan untuk suku kata /ka/, /ki/, /ku/, /kə/, /ke/ dan

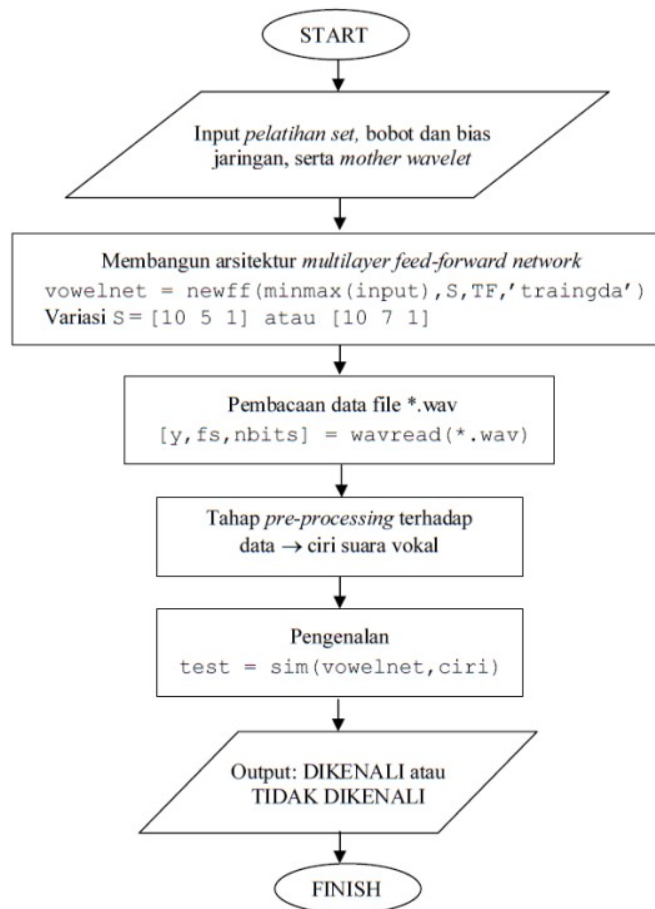
/ko/ targetnya berturut-turut adalah 2, 4, 6, 8, 10 dan 12.

Proses pelatihan akan terus dilakukan hingga *output* jaringan mencapai nilai *minimum error (goal)* sebesar 0.0001 terhadap target yang telah ditetapkan. Selama proses pelatihan berlangsung, nilai bobot dan bias awal jaringan yang awalnya adalah berupa nilai acak dengan *range* antar 0 hingga 1 atau -1 selalu

diperbarui untuk setiap iterasi agar nilai *error* jaringan dapat terus dikurangi. Satu kali proses *backpropagation* dihitung sebagai satu iterasi. Jumlah iterasi yang dilakukan hingga *minimum error* berhasil dicapai ditunjukkan oleh nilai *epoch*. Setelah nilai *minimum error* mampu dicapai maka didapatkan nilai bobot dan bias akhir jaringan yang kemudian digunakan untuk pengujian jaringan.



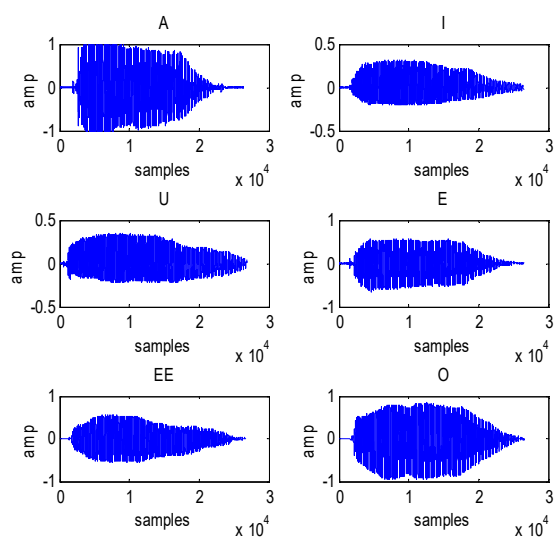
Gambar 4. Flowchart pelatihan JST



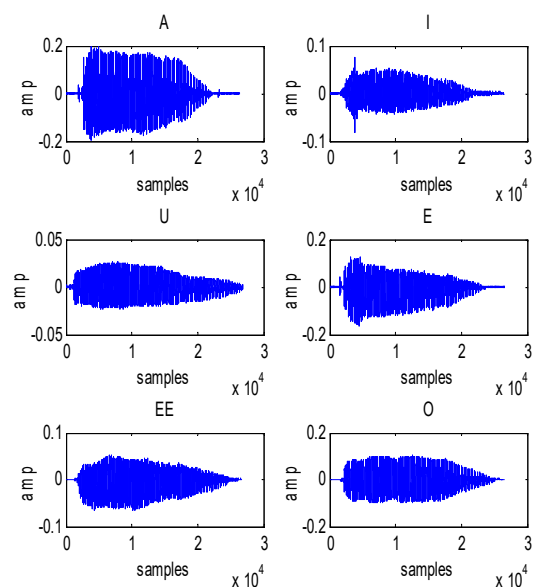
Gambar 5. Flowchart pengujian JST untuk pengenalan suara vokal

HASIL DAN PEMBAHASAN

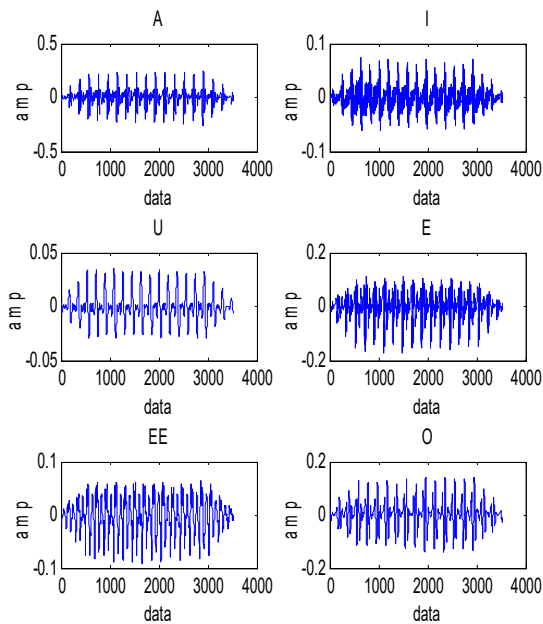
Bentuk Sinyal Suara Vokal



Gambar 6. Sinyal suara vokal, $f_s = 44100$ Hz



Gambar 7. Sinyal suara vokal yang telah di *pre-emphasize filter*



Gambar 8. Sinyal suara vokal yang telah di *frame blocking* dan *windowing*

Hasil Ekstraksi Ciri

Dekomposisi TWD terhadap setiap suara vokal menghasilkan koefisien-koefisien wavelet yang terdiri dari koefisien aproksimasi (cA) dan koefisien detail (cD). Nilai minimum, maksimum, *mean* dan standar deviasi dari koefisien-koefisien wavelet tersebut diekstraksi sebagai ciri suara vokal, sehingga ciri suara vokal yang didapat adalah:

Dekomposisi level 2

Ciri suara vokal berjumlah 12, terdiri dari nilai min cD1, min cD2, max D1, maksimum cD2, *mean* cD1, *mean* cD2, std cD1, std cD2, min cA, max cA, *mean* cA dan std cA.

Dekomposisi level 4

Ciri suara vokal berjumlah 20, terdiri dari nilai min cD1, min cD2, min

cD3, min cD4, max cD1, max cD2, max cD3, max cD4, *mean* cD1, *mean* cD2, *mean* cD3, *mean* cD4, std cD1, std cD2, std cD3, std cD4, min cA, max cA, *mean* cA dan std cA.

Dekomposisi level 6

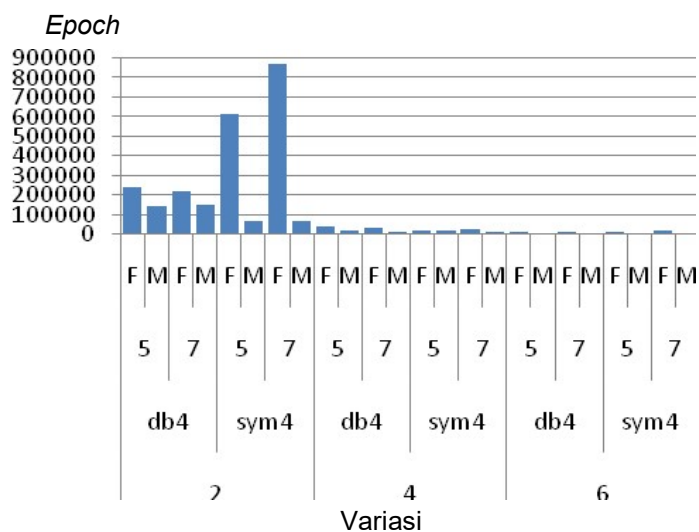
Ciri suara vokal berjumlah 28, terdiri dari nilai min cD1, min cD2, min cD3, min cD4, min cD5, min cD6, max cD1, max cD2, max cD3, max cD4, max cD5, max cD6, *mean* cD1, *mean* cD2, *mean* cD3, *mean* cD4, *mean* cD5, *mean* cD6, std cD1, std cD2, std cD3, std cD4, std cD5, std cD6, min cA, max cA, *mean* cA dan std cA.

Hasil Pelatihan Jaringan

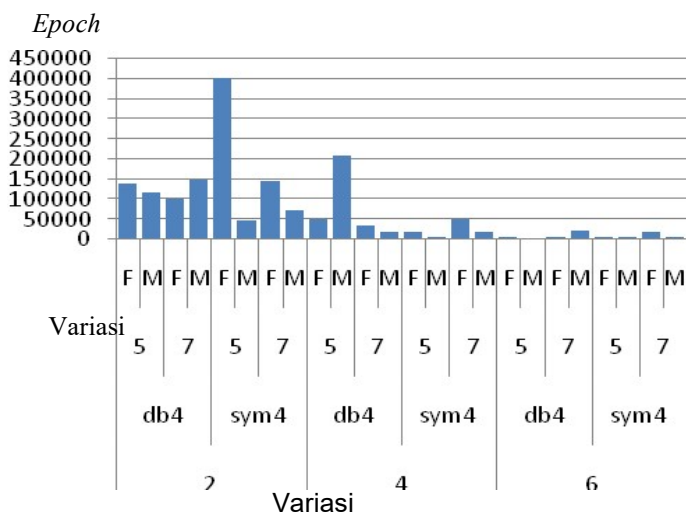
Gambar 9 dan **Gambar 10** menunjukkan grafik perbandingan *epoch* pelatihan jaringan. Pelatihan ini menghasilkan nilai bobot dan bias akhir jaringan tersebut.

Hasil Pengujian Jaringan

Penentuan keputusan dikenali atau tidaknya suara vokal tertentu dengan cara membandingkan *output* dan *threshold range* dari target jaringan. Nilai *threshold* yang ditetapkan adalah 0.8. Hasil pengujian terhadap suara vokal untuk data uji pertama dengan variasi *mother wavelet db4*, level dekomposisi 2 dan jumlah *neuron* pada *hidden layer* kedua adalah 5 ditunjukkan oleh **Gambar 11**.



Gambar 9. Grafik perbandingan nilai epoch pelatihan jaringan untuk suara vokal setiap variasi (F = perempuan, M = laki-laki)



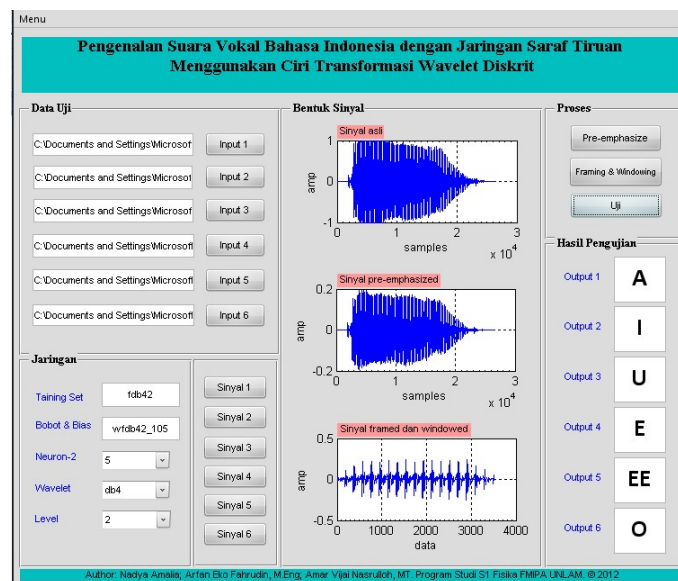
Gambar 10. Grafik perbandingan nilai epoch pelatihan jaringan untuk suku kata setiap variasi (F = perempuan, M = laki-laki)

Tabel 2. Penentuan keputusan dikenali atau tidaknya suara vokal

Output Jaringan	Keputusan	Output
$0.2 \leq test \leq 1.8$	dikenali sebagai /a/	A
$2.2 \leq test \leq 3.8$	dikenali sebagai /i/	I
$4.2 \leq test \leq 5.8$	dikenali sebagai /u/	U
$6.2 \leq test \leq 7.8$	dikenali sebagai /e/	E
$8.2 \leq test \leq 9.8$	dikenali sebagai /ə/	EE
$10.2 \leq test \leq 11.8$	dikenali sebagai /o/	O
lainnya	tidak dikenali	X

Tabel 3. Penentuan keputusan dikenali atau tidaknya suku kata

<i>Output Jaringan</i>	<i>Keputusan</i>	<i>Output</i>
$1.2 \leq test \leq 2.8$	dikenali sebagai /ka/	KA
$3.2 \leq test \leq 4.8$	dikenali sebagai /ki/	KI
$5.2 \leq test \leq 6.8$	dikenali sebagai /ku/	KU
$7.2 \leq test \leq 8.8$	dikenali sebagai /ke/	KE
$9.2 \leq test \leq 10.8$	dikenali sebagai /kə/	KEE
$11.2 \leq test \leq 12.8$	dikenali sebagai /ko/	KO
lainnya	tidak dikenali	X

Gambar 11. Hasil pengujian suara vokal untuk data uji pertama *wfdb42_105*

Penentuan *threshold* tersebut tergantung selisih nilai target dari *input* ciri suara vokal yang satu terhadap target dari *input* ciri suara vokal yang lain. Semakin besar selisih nilai target-target tersebut, semakin besar pula nilai *threshold* yang dapat digunakan.

Berdasarkan hasil pengujian jaringan terhadap data uji, terdapat suara vokal yang dikenali sebagai vokal yang berbeda, yakni suara vokal /u/ yang dikenali sebagai vokal /o/ atau sebaliknya. Hal ini kemungkinan

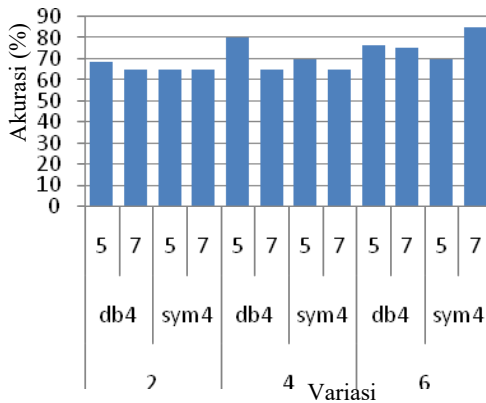
disebabkan oleh pelafalan keduanya sama-sama dengan bibir dibentuk bundar dan menggunakan bagian lidah yang sama yakni bagian belakang.

Tingkat pengenalan suara vokal bahasa Indonesia ditunjukkan oleh akurasi hasil pengujian jaringan terhadap data uji, di mana:

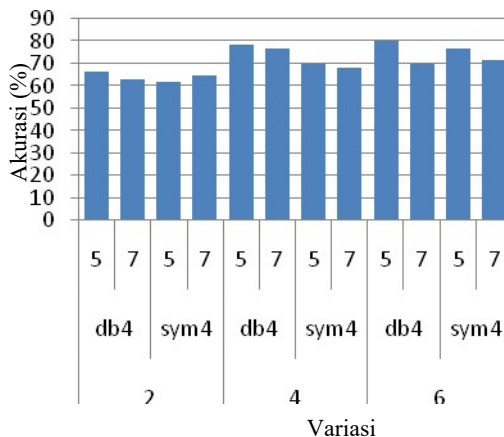
$$Akurasi = \frac{\text{Data benar}}{\text{Total data}} \times 100\% \quad \dots(10)$$

Data benar merupakan jumlah data suara vokal ataupun suku kata yang berhasil dikenali dengan benar,

sementara data salah ada jumlah suara vokal yang mengalami kesalahan pengenalan (dikenali sebagai suara vokal yang lain) ataupun tidak dikenali sama sekali dan total data menunjukkan jumlah data keseluruhan, dalam hal ini adalah 60.



Gambar 12. Grafik perbandingan tingkat pengenalan suara vokal



Gambar 13. Grafik perbandingan tingkat pengenalan untuk suku kata

Selain faktor-faktor yang terkait sistem di dalam JST, tingkat pengenalan sebenarnya juga dipengaruhi oleh faktor fisis sampel suara vokal itu sendiri. Adanya derau pada rekaman sampel suara menghasilkan ciri yang tidak hanya

mencirikan sinyal asli sehingga menghasilkan nilai koefisien aproksimasi yang berbeda-beda. Perekaman sinyal suara vokal yang tepat juga mempunyai pengaruh. Cara perekaman yang baik adalah ucapan diucapkan sewajar mungkin dan tidak dibuat-buat, suara diucapkan tidak terlalu keras dan tidak terlalu lemah.

KESIMPULAN

1. Ciri suara vokal bahasa Indonesia melalui TWD, yaitu: dekomposisi level 2 menghasilkan 12 ciri, dekomposisi level 4 menghasilkan 20 ciri dan dekomposisi level 6 menghasilkan 28 ciri.
2. Proses pelatihan paling cepat dilakukan untuk ciri hasil dekomposisi level 6. Sementara proses pelatihan paling lambat dilakukan untuk ciri hasil dekomposisi level
3. Pengenalan dengan variasi *mother wavelet sym4*, level dekomposisi 6 dan jumlah *neuron* pada *hidden layer* kedua adalah 7 menghasilkan tingkat pengenalan paling baik yaitu 85%, sementara untuk suku kata adalah dengan variasi *mother wavelet db4*, level dekomposisi 6 dan jumlah *neuron* pada *hidden layer* kedua adalah 5 yaitu 80%.

DAFTAR PUSTAKA

- [1] [PPPBDEPDIKBUD] Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan Indonesia. 1996. *Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan dan Pedoman Umum Pembentukan Istilah*. CV Pustaka Setia: Bandung.
- [2] Sunny, S., David P. S. Dan K. P. Jacob. 2012. Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words. *International Journal of Computer Applications*, Vol. 38 No. 9.
- [3] Shukla, P.. 2003. Complex Wavelet Transforms and Their Applications. *Thesis*, University of Strathclyde: Glasgow.
- [4] Agustini, K.. 2007. Biometrik Suara dengan Transformasi Wavelet Berbasis Orthogonal Daubechies. *50 Gematek Jurnal Teknik Komputer, Volume 9 Nomor 2*.
- [5] Lalipola, H.. 1981. A Generative Approach to The Phonology of Bahasa Indonesia. *Pasific Linguistics* 34.
- [6] Mallat, S. G.. 1989. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7.
- [7] Addison, P. S.. 2002. *The Illustrated Wavelet Transform Handbook*. IOP Publishing Ltd.
- [8] Merry, R. J. E.. 2005. Wavelet Theory and Applications. *Thesis*, Eindhoven University of Technology: Eindhoven.
- [9] Strang, G. dan T. Nguyen. 1997. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Second Edition.
- [10] Ambarjati, M., B. L. Widjiantoro, dan A. Rahamadasah. 2011. Identifikasi Suara Ucapan Berbasis Fitur Non-akustik dengan Menggunakan Jaringan Syaraf Tiruan. *Undergraduate Paper*, ITS: Surabaya.
- [11] Schneiders, M. G. E.. 2001. Wavelets In Control Engineering. *Master's Thesis*, Eindhoven University of Technology. DCT nr. 2001.38.
- [12] Yana, E.. 2005. Pengantar Jaringan Saraf Tiruan. *Artikel kuliah*. http://trirezqariantoro.files.wordpress.com/2007/05/jaringan_saraf_tiruan.pdf
- [13] Puspitaningrum, D.. 2006. *Pengantar Jaringan Saraf Tiruan*. Penerbit Andi: Jogjakarta.
- [14] Chaudari, J. C.. 2010. Design of Artificial Back Propagation Neural Network for Drug Pattern Recognition. *International Journal on Computer Science and Engineering*.