IDENTIFIKASI PEMBICARA MENGGUNAKAN JARINGAN SARAF TIRUAN PROBABILISTIK

JAYANTA[1)]
[1)]Staf Pengajar Fakultas Ilmu Komputer
UPN "Veteran" Jakarta

ABSTRAK

*Tulisan ini merupakan penjelasan hasil penelitian pada bidang sistem identifikasi pembicara menggunakan Jaringan Saraf Tiruan Probabilistik (JSTP) pada komputer personal secara off line. Suara sebagai salah satu data biometrik digunakan sebagai bahan baku penelitian, dicuplik pada frekuensi 16000 Hz dengan ketelitian 16 bit dan durasi rekam suara 2 detik. Sebanyak 10 orang dewasa berusia 21 tahun hingga 46 tahun dilibatkan dalam penelitian ini untuk mengucapkan kata sandi **"Sembilan"**. Ciri suara didapat melalui proses ekstraksi ciri dengan teknik mel-frequency cepstral coefficient (MFCC) dengan parameter koefisien mel 16 dan 20, utuk memudahkan proses ekstraksi ciri digunakan lebar waktu frame 40 ms dan overlap 50% dari lebar waktu frame. Sebanyak N-1 data ciri digunakan sebagai data pelatihan untuk merekonstruksi model JST Probabilsitik, dan data ciri sisanya dijadikan data pengujian model JST Probalisitik. Hasil pengujian memperlihatkan, bahwa keakuratan JSTP dan MFCC16 mengidentifikasi suara pembicara mencapai 94 %, sedangkan untuk JSTP dan MFCC20 keakuratan identifikaasi mencapai 96 %.*

***kata kunci** : jaringan saraf tiruan probabilistic(JSTP), MFCC, sistem identifikasi pembicara.*

A.PENDAHULUAN

Suara merupakan sumber data alamiah yang dimiliki manusia dan dapat memberikan banyak sekali informasi, antara lain : informasi mengenai rangkaian huruf pembentuk kata atau kalimat ; informasi bahasa yang di gunakan untuk berbicara; emosi; jenis kelamin; usia serta informasi identitas pemilik suara.

Salah satu parameter manusia yang dikenal dengan data biometrik, dan mempunyai keunggulan sifat tidak dapat dihilangkan, dilupakan, atau dipindahkan dari satu orang ke orang lain. Di masa depan, teknologi biometrik akan mirip fenomena komputer yang kemudian menjadi bagian dari sebuah kebutuhan hidup sehari-hari.

Contoh dari penerapan aplikasi teknologi biometrik, adalah penerapan sistem pengenalan suara pada aplikasi layanan nasabah perbankan melalui jaringan telepon, dimana penggunaan suara dilakukan untuk mengakses layanan yang disediakan pihak perbankan.

Sistem identifikasi pembicara tidak lain adalah bagian dari sistem pengenalan suara, dimana sistem bekerja dengan menangkap sinyal suara, kemudian dianalisis untuk menemukan ciri penting setiap suara yang akan diklasifikasikan

kedalam kelompok ciri suara yang telah tersimpan dalam basis data.

Produk sistem pengenalan suara, secara ekonomi dapat memberikan nilai jual yang sangat menjanjikan. Penjualan produk teknologi informasi berbasis suara sistem pengenalan suara) pada tahun 1997, memberikan nilai sebesar 500 juta dolar Amerika, dan meningkat menjadi 38 milyar dolar Amerika pada tahun 2003. Selain keuntungan finansial, sistem juga dapat diterapkan melalui jaringan telepon tetap maupun selular.

Mendapatkan ciri dari setiap suara merupakan masalah yang harus dihadapi ketika kita akan mngembangkan sistem pengenalan suara atau sistem identifikasi pembicara, mendapatkan ciri penting dari setiap suara, akan mempermudah proses identifikasi suara.

MFCC merupakan salah satu teknik yang bisa digunakan untuk mendapatkan ciri penting dari setiap suara. Penelitian ini, menggunakan koefisien mel 16 dan 20 pada penerapan metode MFCC, disebut dengan MFCC16 dan MFCC20. Untuk memudahkan penerapan metode MFCC, sinyal suara dijital dipecah menjadi beberapa *frame* dengan panjang setiap *frame* sebesar 40 ms dengan jarak antar *frame* 20 ms.

Untuk memudahkan penelitian digunakan pendekatan *text-dependent*, dimana kata yang akan diucapakan telah disiapkan sebelumnya, dan pembicara tinggal mengucapkan kata tersebut.

Kata sandi "*Sembilan*" disiapkan untuk diucapkan oleh 10 orang relawan, setiap relawan diminta mengucapkan kata tersebut sebanyak 10 kali, dengan demikian akan terkumpul sebanyak 100 sinyal suara dijital.

Sembilan puluh sample data suara digunakan sebagai data pelatihan dalam merekonstruksi jaringan saraf tiruan probabilistik, terbagi kedalam sepuluh kelas untuk sepuluh pembicara berbeda.

Setelah dilakukan proses pelatihan, sistem diuji coba untuk mengidentifikasi suara pembicara menggunakan data yang sebelumnya telah direkam pada basisdata suara (pengujian dilakukan secara *offline*).

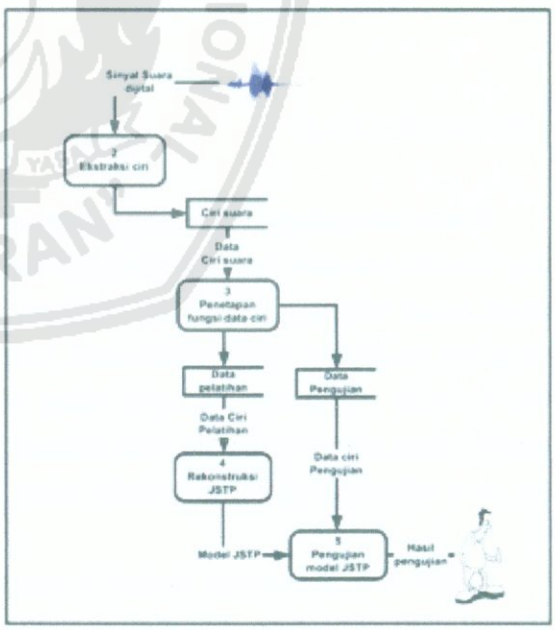
Pengujian sistem memperlihatkan tingkat keberhasilan identifikasi sebesar 96 %.

B. SISTEM IDENTIFIKASI PEMBICARA

Setelah proses pengumpulan suara, alir kegiatan penelitian dibidang sistem identifikasi pembicara digambarkan dengan blok diagram alir data, seperti disajikan pada Gambar 1:

Garis besar kegiatan penelitian terbagi menjadi 4 proses, yaitu:

- proses pengumpulan suara;
- Proses Ekstraksi ciri suara;
- Rekonstruksi dan pelatihan model JSTP;
- Pengujian sistem, divalidasi dengan metode *leave-one out*.



Gambar1 Sistem identifikasi pembicara

a. Proses Pengumpulan Suara

Suara dikumpulkan menggunakan perangkat mikrofon dan kartu suara standar PC yang ditancapkan pada

komputer personal dengan prosesor AMD Athlon 1700+ dan memori utama 256 MB.

Pencuplik suara dilakukan pada frekuensi 16000 Hz dengan ketelitian 16 bit (2 byte) dan durasi perekaman 2 detik untuk satu kali pengucapan kata sandi "***Sembilan***". Dasar pemilihan frekuensi pencuplikan suara adalah asumsi bahwa sinyal percakapan berada pada daerah frekuensi 300-3400 Hz sehingga pencuplikan memenuhi kriteria Nyquist yang menyatakan :

$f_S \geq 2xf_h \quad f_h = f_{in}{}^{tertinggi}$

Dari 10 orang relawan, setiap orang diminta untuk melakukan pengucapan kata sandi sebanyak 10 kali, sehingga dari pengucapan kata sandi tersebut terkumpul 100 suara. Setiap suara berbentuk vector dengan berdimensi 32000.

b. Ekstraksi Ciri

Proses ekstraksi ciri dilakukan untuk mendapatkan nilai ciri dari setiap suara yang diamati. Dengan menerapkan metode MFCC diharapkan akan didapat serangkaian ciri suara yang terbaik untuk digunakan proses identifikasi pembicara. Langkah pada proses ekstraksi ciri, adalah:

- Meratakan spektral sinyal suara yang telah direkam (***Preemphasis***). Pada langkah ini, sinyal suara dijital disaring menggunakan metode penyaringan *FIR* orde satu

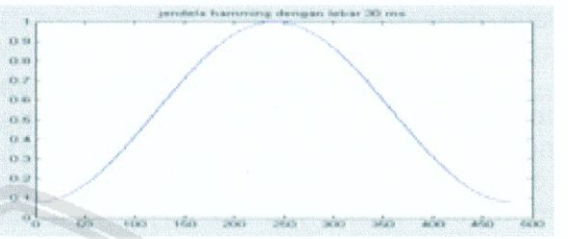
  $\tilde{s}(n) = s(n) - \tilde{a}s(n-1)$

- ***Frame Blocking.*** Pada tahap ini sinyal suara yang telah teremphasi dipecah menjadi beberapa *frame* dengan masing-masing *frame* memuat *N* nilai data suara dan *frame-frame* yang berdekatan dipisahkan sejauh *M* nilai data suara (*overlap*).

- ***Windowing.*** Pada langkah ini dilakukan pembentukan fungsi *window hamming* menggunakan persamaan,

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

  Hasil eksekusi persamaan diatas, disajikan pada Gambar 2.



Gambar 2 bentuk *window hamming*

Windowing atau pembobotan *window* dilakukan terhadap setiap *frame* yang telah dibentuk pada langkah sebelumnya.

- ***Fast Fourier Transform* (FFT).** Proses *Fast Fourier Transform* (FFT) ini dilakukan setelah didapat hasil pembobotan *window*.FFT akan mentransformasikan *frame-frame* hasil pembobotan *window hamming* kedalam domain frekuensi (hertz).
- Setiap *frame* yang telah melalui proses FFT disaring dengan *mel-triangular filter bank* 16 dan 20, kemudian diarahkan menjadi frekuensi mel berdasarkan skala mel yang didapat menggunakan formula:

  $M(f) = 2595.log_{10}\left(1 + \frac{f}{700}\right)$

  dimana *f* adalah frekuensi dalam hertz.
- Frekuensi mel kemudian dikompresi menggunakan fungsi **log** sebelum ditransformasikan kembali kedalam domain frekuensi (hertz) dengan fungsi DCT.

c. Jaringan Saraf Tiruan Probabilistik

Jaringan saraf tiruan probabilistik (JSTP) atau *probabilistic neural networks* (PNN), diperkenalkan pertama kali oleh D.F Specht pada tahun 1988, sebagai

jaringan saraf tiruan dengan 3 lapisan tersembunyi setelah *input layer*, yaitu: *pattern layer, summation layer, output layer*, bersifat *feed-forward*, dieksekusi hanya dengan satu kali proses (*one pass*).

Keuntungan yang diberikan JSTP adalah kemudahan untuk memodifikasi jaringan, ketika dilakukan penambahan atau pengurang kelas atau data pelatihan yang digunakan.

Sedangkan kelemahan JSTP adalah peningkatan dari penggunaan ruang memori komputer, dan waktu komputasi, ketika data pelatihan yang digunakan bertambah besar, hal ini terjadi karena semua data pelatihan harus dimasukkan ke dalam algoritma JSTP.

Kerja JST Probabilistik, didasarkan pada penghitungan nilai fungsi kepekatan peluang *(fi(x))* untuk setiap data (vektor). Fungsi *(fi(x))* merupakan fungsi pengambilan keputusan *Bayes (gi(x))*, untuk data (vektor) *x* dan *xij* yang telah dinormalisasi. Persamaan fungsi *fi(x)* atau *gi(x)*, tuliskan sebagai berikut.

$$f_i(x) = \frac{1}{(2\pi)^{\frac{\rho}{2}} \sigma^{\rho} M_i} \sum_{j=1}^{M_i} \left[\exp\left(- \frac{((x - x_{ij})^T . (x - x_{ij}))}{2.\sigma^2} \right) \right]$$

dimana:

T : Transpose
I : Jumlah kelas
J : Jumlah pola
Xij : Vektor pelatihan ke j dari kelas i
X : Vektor pengujian
Mi : Jumlah vektor pelatihan dari kelas ***i***
ρ : Dimensi vektor x
σ : Faktor penghalus

Sampel data untuk data pelatihan tidak sama dengan sampel data untuk data pengujian JSTP. Diagram arsitektur JSTP, disajikan pada Gambar 3.

Posisi node-node setelah lapisan input, adalah:

1 Node Pattern Layer. Untuk setiap kelas digunakan 1 node pola yang berisi sejumlah data pelatihan. Setiap node pola, merupakan perkalian titik (*dot product*) dari vektor masukkan x yang akan diklasifikasikan, dengan vektor bobot x_{ij}, yaitu $Z_i = x \; . \; x_{ij}$, kemudian di lakukan operasi non-linier terhadap Z_i sebelum menjadi keluaran yang akan mengaktifkan lapisan berikutnya, yaitu lapisan penjumlahan (summation layer). Persamaan $exp[(Zi - 1)/\sigma^2]$ digunakan untuk melakukan operasi non-linier, dan jika vektor x dan x_{ij}, dinormalisasikan terhadap panjang vektor, maka persamaan yang digunakan pada lapisan pola (*pattern layer*), adalah:

$$exp\left[- \frac{(x - x_{ij})^T (x - x_{ij})}{2\sigma^2} \right]$$

2 Node *Summation Layer*, akan menerima masukkan dari node lapisan pola yang terkait dengan kelas yang ada, persamaan yang digunakan pada lapisan ini, adalah:

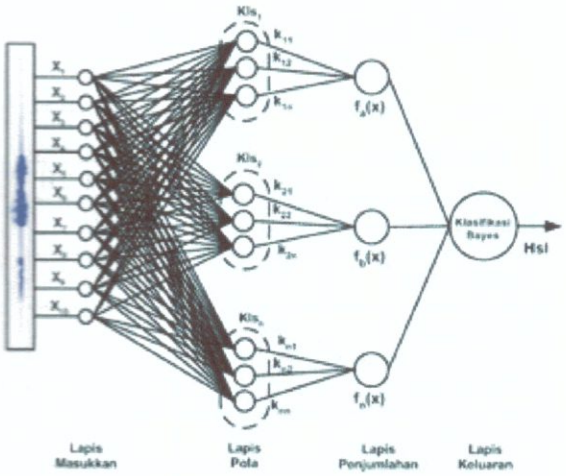
$$\sum_{i=1}^{N} exp\left[- \frac{(x - x_{ij})^T (x - x_{ij})}{2\sigma^2} \right]$$

3 Node lapisan Keluaran (*Output Layer*), menghasilkan keluaran biner (0,1), dan hanya mempunyai variabel bobot tunggal *Ck*. *Ck* dihitung mengguna-kan persamaan:

$$C_k = - \frac{h_{jk} \; l_{jk}}{h_{ik} \; l_{ik}} \cdot \frac{m_{ik}}{m_{jk}}$$

dimana:

m_{ik} = Jumlah pelatihan pola dari kelas θ_{ik};
m_{jk} = Jumlah pelatihan pola dari kelas θ_{jk}



Gabar 3 Arsitektur Jaringan Saraf Tiruan Probabilistik

C. HASIL-HASIL PERCOBAAN

a. Rekonstruksi Struktur Jaringan Saraf Tiruan Probabilistik

Jumlah kelas yang diunakan ketika dilakukan rekonstruksi JST Probabilistik menggambarkan banyaknya pembicara yang dilibatkan dalam penelitian ini. Banyak data ciri dari setiap pembicara untuk dijadikan data pelatihan dan data acuan ditempatkan pada setiap kelas, digunakan oleh JSTP sebagai data acuan untuk mengetahui suara yang diterima akan teridentifikasi sebagai suara siapa. Kemudian untuk mengarahkan hasil identifikasi digunakan nilai faktor penghalus (ó), pada penelitian ini besar nilai ó yang digunakan adalah 9. Langkah tersebut diatas merupakan tahapan dalam merekonstruksi JSTP.

b. Ekstraksi ciri

Proses *frame blocking* yang dilakukan pada sistem ini ditetapkan tiap 40 mili detik dengan jarak antar *frame* adalah 50% dari lebar waktu *frame*. Jadi dengan kecepatan cuplik sebesar 16000 Hz maka tiap *frame* akan berisi 640 *byte* data dengan jarak antar *frame* 320 *byte* data atau dengan kata lain *overlap* yang terbentuk sebesar 320 *byte* data. Dengan ketentuan *frame* seperti di atas, maka untuk setiap suara yang digunakan sebagai bahan penelitian akan terbentuk

$$\frac{(16000-640)}{(640-320)}+1=49 \text{ buah frame.}$$

Proses ekstraksi ciri diterapkan pada 49 *frame* suara. Penerapan metode MFCC16 akan mereduksi jumlah data pada setiap *frame* dari 640 menjadi 99, sedangkan jumlah *frame* yang ada akan tereduksi menjadi 15 buah frame, dengan demikian untuk setiap suara yang digunakan akan dihasilkan 15 * 99 = 1485 *byte* data ciri. Pada penerapan metode MFCC20 akan mereduksi jumlah *frame* menjadi 19, dengan demikian untuk setiap suara akan dihasilkan 19*99 = 1881 *byte* data ciri.

c. Pengujian Sistem

Pada tahap awal, uji identifikasi dilakukan terhadap sinyal suara yang sama persis dengan yang telah dilatihkan (*training data set*) dan didapat hasil bahwa error yang terjadi sebesar 0 % atau dengan kata lain tingkat akurasi sistem untuk mengenali pola *training data set* mencapai 100 % (Tabel 1).

Pengujian dilajutkan menggunakan basisdata suara yang telah disiapkan sebagai data pengujian, dimana data ini berbeda dari data yang disiapkan sebagai data pelatihan.

Proses pengujian menghasilkan rata-rata salah identifikasi sebesar 6 % untuk data yang diproses dengan MFCC16, dengan kata lain sistem mampu mengidentifikasi suara pembicara dengan akuratan 94 %. Sedangkan untuk data yang diproses dengan MFCC20 menghasilkan rata-rata salah identifikasi sebesar 4 % atau dengan kata lain keakuratan sistem untuk mengidentifikasi pola suara meningkat sebesar 2 %, yaitu: dari 94 % menjadi 96% (Tabel 2).

Kesalahan identifikasi disumbangan oleh suara pembicara 2, 3, 7 dan 9, untuk data yang proses menggunakan metode MFCC16. Sedangkan untuk data ciri yang proses dengan MFCC20, kesalahan

identifikasi terjadi hanya pada pembicara 2 dan 3 (Tabel 2).

Data ciri pembicara 3 menghasilkan kesalahan lebih tinggi dibandingkan data ciri pembicara 2, 7 dan 9. Meski terjadi kesalahan, namun keakuratan sistem dalam mengidenifikasi setiap suara pembicara masih berada diatas 95% (Tabel 2).

TABEL 1. ERROR RATE PADA PENGUJIAN DENGAN TRAINING DATA SET

Pembicara	Error rate	
	MFCC16	MFCC20
Pembicara 1	0%	0%
Pembicara 2	0%	0%
Pembicara 3	0%	0%
Pembicara 4	0%	0%
Pembicara 5	0%	0%
Pembicara 6	0%	0%
Pembicara 7	0%	0%
Pembicara 8	0%	0%
Pembicara 9	0%	0%
Pembicara 10	0%	0%
Error rata-rata	0%	0%

TABEL 2. ERROR RATE PADA PENGUJIAN DENGAN TESTING DATA SET

Pembicara	Error rate	
	MFCC16	MFCC20
Pembicara 1	0%	0%
Pembicara 2	10%	10%
Pembicara 3	30%	30%
Pembicara 4	0%	0%
Pembicara 5	0%	0%
Pembicara 6	0%	0%
Pembicara 7	10%	0%
Pembicara 8	0%	0%
Pembicara 9	10%	0%
Pembicara 10	0%	0%
Error rata-rata	6%	4%

D. KESIMPULAN

Dari hasil percobaan dapat disimpulkan, bahwa: untuk lebar *frame* 40 ms dengan jarak antar *frame* 20 ms, peningkatan nilai koefisen mel dari 16 ke 20 dapat peningkatan akurasi indentifikasi suara pembicara hingga 2 %.

DAFTAR PUSTAKA

Campbell, J.P., 1997, Speaker Recognition: A Tutorial, *Proceedings of the IEEE*, Vol. 85, No. 9.

Specht D F, Shapiro P D. 1991. Generalization Accuracy of Probabilistic Neural Networks Compared with Back-Propagation Networks. *IEEE Transaction on Neural Networks* 1:887-892

Furui S. 1997. Recent advances in speaker recognition. *Pattern Recognition Letters* 18: 859 - 872.

Ganchev T. D, 2005, "Speaker Recognition" [desertasi] Department of Computer and Electrical Engineering University of Patras, Yunani. http://www.wcl.ee.upatras.gr/ai/papers/Ganchev_PhDThesis.PDF

Rabah Y. 2004, Speech Recognition. www.earlham.edu/~rabahyo/survey.doc.

Rabiner L, Juang BH. 1993. *Fundamental of Speech Recognition*. New Jersey: PTR Prentice-Hall, Inc.