SISTEM PENGENALAN WICARA BERDASARKAN *CEPSTRUM* DAN *HIDDEN MARKOV MODEL*

Ivanna K. Timotius, Danie Kurniawan

Fakultas Teknik Elektronika dan Komputer, Program Studi Teknik Elektro,
Universitas Kristen Satya Wacana, Salatiga, Indonesia
ivanna_timotius@yahoo.com, danie.kurniawan@epcos.com

Intisari

Sistem pengenal wicara merupakan sistem yang bisa mengerti kata-kata yang diucapkan manusia. Sistem pengenal wicara dapat dimanfaatkan di berbagai bidang kehidupan. Tulisan ini bertujuan untuk merealisasikan sistem pengenal wicara dengan menggunakan metode cepstrum dan *Hidden Markov Model*. Sistem ini diterapkan untuk mengenali 23 kata yang diperoleh dengan mencuplik sinyal dari mikrofon yang terhubung ke kartu suara pada komputer. Dari hasil pengujian didapatkan tingkat keberhasilan pengenalan 76,52%

Kata kunci: pengenal wicara, cepstrum, *hidden Markov model*.

1. Pendahuluan

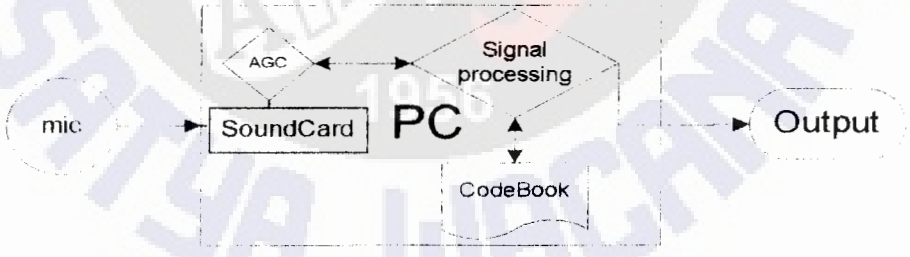
Sistem pengenal wicara (*speech recognizer*) merupakan sistem yang bisa mengerti kata-kata (tanpa perlu mengetahui arti dari kata sebenarnya) yang diucapkan manusia dan merespon terhadap kata yang didengarnya tersebut. Suatu sistem pengenal wicara dapat dimanfaatkan sebagai *voice dialer* pada telepon selular, pengubah sinyal wicara ke bentuk teks, sistem identifikasi pada sistem keamanan, fasilitas pada suatu *smart home*, alat bantu tuna daksa, dan lain sebagainya.

Kesulitan yang sering dihadapi oleh sistem pengenal wicara adalah terpengaruhnya sinyal wicara keadaan pembicara saat berlangsungnya proses pengucapan (meskipun sinyal wicara diucapkan oleh pembicara yang sama). Misalnya *pitch* yang berbeda karena perbedaan suasana hati pembicara (sedih, marah, senang), durasi,

kecepatan dan intonasi suara, maupun bentuk efek yang ditimbulkan oleh posisi rongga hidung ketika sinyal suara diciptakan dan derau lingkungan [1]. Selain itu, orang yang berbeda akan menghasilkan bentuk sinyal wicara yang berbeda juga.

Tulisan ini mengembangkan suatu sistem pengenal wicara berdasarkan cepstrum dan *Hidden Markov Model* (HMM). Penggunaan cepstrum dalam analisis dan pemrosesan sinyal wicara telah diteliti dan dikembangkan cukup lama [2]. Metode cepstrum yang dapat digunakan sebagai metode dekonvolusi sinyal telah diketahui secara matematis mampu dipergunakan sebagai fitur pengolahan sinyal wicara yang cukup baik [1]. Sebagai klasifikator digunakan HMM. HMM dapat dipandang sebagai pernyataan *state machine* yang *state*-nya berpindah – pindah setiap waktu tertentu. Perpindahan antar *state* dalam sebuah HMM berdasarkan probabilitas tertentu [3].

Sistem yang dikembangkan dapat mengambil sinyal masukan melalui mikrofon melalui kartu suara PC, mengolah sinyal suara tersebut untuk diambil fitur, kemudian mengolahnya untuk proses pengenalan. Blok diagram dari sistem ditunjukkan oleh gambar 1. *Automatic Gain Control* (AGC) diterapkan untuk mengkompensasi penguatan sinyal wicara supaya pada data pencuplikan sinyal wicara tidak terjadi pemotongan bit hasil kuantisasi sistem. Proses di dalam sistem terdiri dari pemrosesan awal, ekstraksi fitur dengan menggunakan metode cepstral dan proses klasifikasi dengan menggunakan HMM.



Gambar 1. Blok diagram sistem.

2. Pemrosesan Awal Sinyal Wicara

Pemroses awal sinyal wicara berisi AGC, pendeteksi awal dan akhir suatu pengucapan kata, serta bagian segmentasi. Bagian pemroses awal berisi AGC yang mengkompensasi proses kuantisasi pada saat proses penyuplikan berlangsung. Hal ini

bertujuan supaya pada saat proses kuantisasi pencuplikan tidak terjadi pemotongan bit. Bagian pendeteksi awal-akhir (*endpoint detection*) suatu pengucapan kata, yang berfungsi menentukan waktu mulai dan berakhirnya suatu kata. Sedangkan segmentasi bertugas membagi sinyal ke dalam bentuk segmen-segmen kecil dengan cara menerapkan fungsi penjendela ke dalam sistem.

Proses mendeteksi adanya suatu pengucapan kata ditentukan berdasarkan perhitungan daya sinyal masukan dibandingkan dengan daya derau latar. Sinyal masukan akan dianggap sebagai suatu kata apabila daya sinyal tersebut lebih besar daripada ρ kali daya derau latar, dengan $2 \geq \rho \geq 1$.

Karakteristik utama sinyal wicara akan memiliki sifat akustik yang cukup stabil bila dilakukan analisis sinyal wicara dalam jangka waktu yang singkat (dengan jangka waktu 20-40 mili detik) [1]. Karena itu, suara manusia digambarkan dalam analisis bentuk pendek (*short-term analysis*), dimana sepotong sinyal digunakan untuk mengekstraksi fitur pada suatu saat. Hal ini dilakukan dengan menerapkan fungsi jendela (*window*) biasanya untuk waktu yang singkat pada sepanjang sinyal. Operasi pembagian sinyal ke dalam interval yang pendek disebut penjendelaan dan potongan yang dihasilkan disebut disebut segmen.

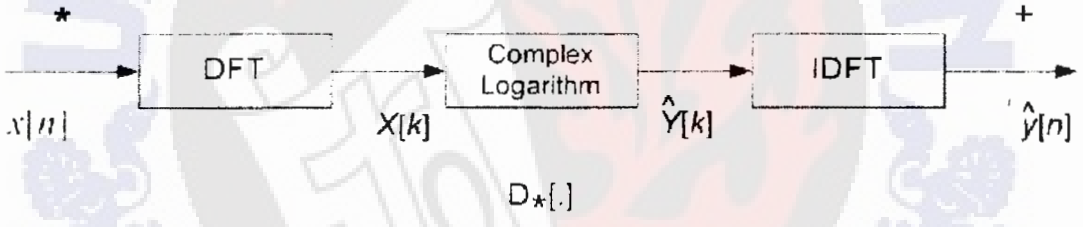
Proses segmentasi dilakukan dengan membagi sinyal wicara ke dalam blok-blok yang kecil dengan menerapkan fungsi penjendela. Proses segmentasi dilakukan tanpa adanya tumpang tindih (*overlapping*) antara segmen yang satu dengan yang lainnya. Setiap segmen akan terdiri dari sejumlah N buah data hasil pencuplikan sinyal $x(n)$. Secara umum, nilai N berkisar pada interval 20-40 ms yang telah terbagi ke bentuk segmen ini kemudian dikalikan dengan fungsi penjendela. Pada tulisan ini dipilih interval 25,6 ms dengan frekuensi cuplik 10 kHz, sehingga N bernilai 256. Fungsi penjendela $w(n)$ yang biasa dipakai adalah jenis jendela Hamming [1][4][5]. Penggunaan jendela Hamming bertujuan untuk menghindari perubahan yang terlalu mendadak pada ujung-ujung tiap segmen dan untuk lebih memberikan efek stasioner pada sinyal.

3. Cepstrum

Sinyal suara manusia $s(n)$ dapat dinyatakan dalam bentuk konvolusi $s(n) = e(n) * h(n)$, dengan $e(n)$ adalah sinyal yang berubah cepat terhadap waktu, sedangkan $h(n)$ adalah sinyal yang berubah lambat terhadap waktu [4]. Jika kedua komponen ini data dipisahkan maka kita dapat mengharapkan pengenalan informasi suara

komponen ini adalah cepstrum. Metode cepstrum merupakan suatu cara dalam proses *homomorphic* untuk menemukan sistem vokal dari *filter* $H(z)$ [2]. Penurosesan sinyal dengan cara ini mengacu pada transformasi ke domain linear dari sinyal yang terkombinasi secara tak linear.

Cepstrum merupakan bentuk transformasi balik dari besaran logaritmik spektral seperti ditunjukkan oleh gambar 2, dengan $x[n]$ merupakan sinyal masukan, $X[k]$ merupakan spektrum sinyal, $\hat{Y}[k]$ merupakan bentuk Logaritmik spektral, dan $\hat{y}[n]$ merupakan cepstral



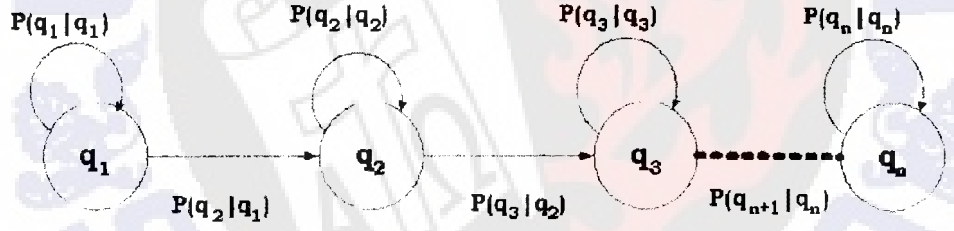
Gambar 2 Cepstrum kompleks.

Penjelasan dari metode ini adalah dengan memindahkan analisis ke ranah frekuensi, berarti mengubah konvolusi menjadi perkalian. Masalahnya di sini adalah bahwa perkalian merupakan operasi yang tidak linier. Sehingga akan dipakai fungsi logaritmik untuk mengubah operasi perkalian menjadi penjumlahan. Hal inilah yang diinginkan, yaitu pemisahan ke dalam komponen penjumlahan. Lalu dilanjutkan dengan operator linier balik DFT yang akan mentransformasikan kedua komponen ini masing-masing terhadap bagian yang berubah cepat dan bagian yang berubah lambat ($e(n)$ dan $h(n)$). Sehingga kedua komponen ini akan terpisah ke dalam bentuk ranah yang baru, yang disebut ranah *Quefrency* [1]. Pada tulisan ini dipakai 15 koefisien cepstral yang pertama (bagian yang berubah lambat $e(n)$) dan kemudian koefisien-koefisien tersebut dikuantisasi sebelum diproses oleh klasifikator.

4. *Hidden Markov Model*

Sebuah HMM adalah sebuah proses acak ganda (*doubly stochastic process*) [3][6] dengan satu proses acak pokok yang tidak dapat diobservasi (*hidden*), tetapi hanya dapat diobservasi melalui himpunan proses acak lainnya yang menghasilkan urutan simbol [illegible] rantai yang berkaitan antara satu dengan yang lain. Model diskret dari pemodelan ini sering disebut juga rantai Markov diskret (*Discrete Markov Chain*) [6]

Model rantai Markov diskret digambarkan pada gambar 3. Bila terdapat N buah *state* terbatas (*finite state*). Setiap vektor fitur hasil pengamatan akan menempati sebuah *state* $Q = \{q_1, q_2, \ldots, q_N\}$ dengan suatu nilai peluang yang tertentu $P(q_j \mid q_i)$. Keputusan untuk berpindah dari satu *state* ke *state* lainnya didasarkan pada fungsi peluang dan kemiripan karakteristik dari fitur tersebut. Keadaan Q_i ini disebut sebagai *state*. Di dalam sebuah *state*, nilai dari suatu sinyal dapat diukur atau dengan kata lain, sinyal dalam kondisi tersebut memiliki properti yang berbeda.



Gambar 3 Model rantai Markov diskret.

Pada tulisan ini dipakai model rantai Markov yang disederhanakan. Semua nilai peluang $P(q_{n+1} \mid q_n)$ ditentukan sama. Kemiripan karakteristik antar *state* dihitung berdasarkan persamaan berikut:

$$\eta = \frac{1}{M} \sum_{m=1}^{m=M} (x_m - y_m)^2 \qquad (1)$$

dengan η adalah nilai kemiripan suatu *state*, M adalah orde vektor dalam sebuah *state*, x adalah *state* pada model hasil pengamatan, dan y adalah *state* pada model dalam basis data.

State akan berada di posisi yang sama apabila vektor pengamatan yang muncul tidak lebih mirip daripada *state* selanjutnya. Batas kemiripan sebuah *state* diberikan melalui pemberian nilai batas toleransi pada fungsi pembanding kemiripan *state*. Bila vektor pengamatan suatu *state* lebih mirip dengan *state* berikutnya maka akan dilakukan transisi pada model tersebut dan memberikan nilai pada model yang melakukan transisi.

paling tepat

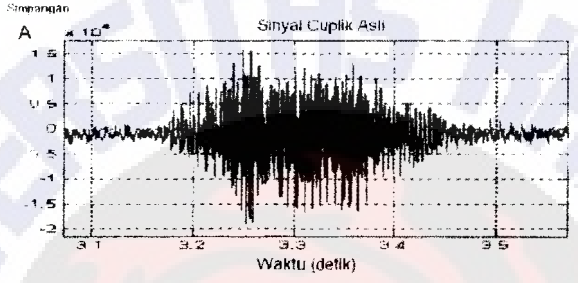
Pada bagian pembelajaran, model Markov yang dihasilkan untuk tiap sinyal wicara disimpan ke dalam basis data. Pada bagian pengenalan, model Markov yang dihasilkan dari sinyal uji akan dibandingkan dengan setiap model kata yang ada pada basis data. Pencarian kata ini berdasarkan pada kemiripan karakteristik *state* yang ada pada model kata. Kemiripan karakteristik diuji dengan menghitung jarak tiap *state* dengan menggunakan persamaan 1. Dari hasil pengujian ini, pada η diberikan suatu batas toleransi kemiripan Δ minimal yang tertentu. Pada skripsi ini diberikan nilai batas toleransi Δ sebesar 2,5. Pemilihan nilai batas toleransi ini didapat dari hasil melakukan percobaan.

5. Implementasi dan Pengujian

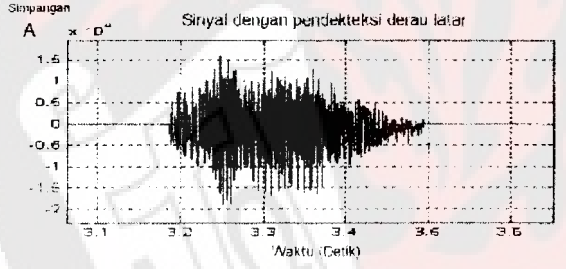
Sistem pengenal wicara dirancang untuk mengenali kata-kata berikut: *Computer, Lamp, Fan, On, Off, Hello, Shutdown, Next, Previous, Stop, One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Zero, Music, Notepad* dan *Email*. Masukan sinyal wicara diperoleh secara langsung lewat *Soundcard* dan dicuplik dengan frekuensi cuplik 10 kHz. Jeda waktu antar kata minimal 10 mili detik. Sistem pengenal yang dibuat telah mampu mengenali kalimat wicara yang terdiri lebih dari dua kata, dengan diberikan batasan jeda waktu antar kata maksimal 2 detik. Apabila jarak jeda antara pengucapan kata pertama dengan kata selanjutnya melebihi batas maksimal dari jeda yang telah ditentukan, maka kata tersebut akan dianggap sebagai bagian dari kalimat yang berbeda. Sistem pengenalan akan dapat berjalan dengan baik bila sinyal lebih besar 20 dB dari *noise* lingkungan.

Gambar 4 menunjukkan pencuplikan sinyal wicara untuk kata "*One*" yang diambil secara langsung melalui mikrofon dengan frekuensi cuplik sebesar 10 kHz. Setelah dilakukan perbandingan dengan daya derau latar didapatkan sinyal seperti

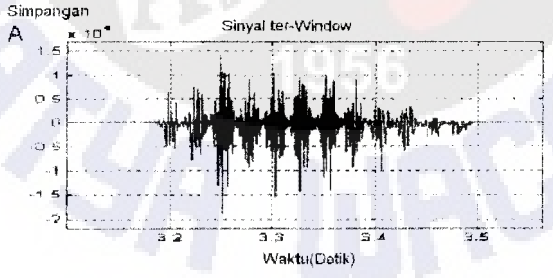
Gambar 5. Setelah ditemukan titik awal suatu pengucapan kata, maka dilakukan segmentasi pada sinyal tersebut dan menerapkan fungsi jendela jenis Hamming sehingga didapatkan sinyal seperti Gambar 6. Setelah sinyal dilewatkan pada fungsi penjendela Hamming, proses dilanjutkan dengan ekstraksi fitur sinyal wicara dengan metode cepstral. Sebagai [illegible] sinyal. Fitur cepstrum hasil ekstraksi yang telah terkuantisasi untuk sinyal wicara "*One*" dapat dilihat pada Gambar 7.



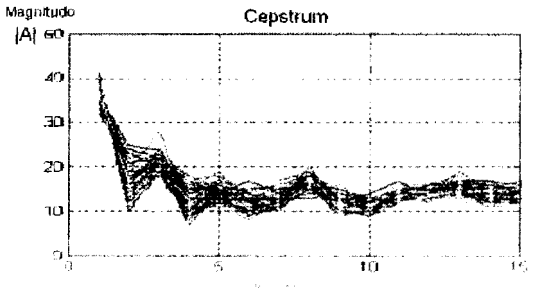
Gambar 4 Sinyal cuplik asli sinyal wicara "*One*"



Gambar 5 Sinyal keluaran pendeteksi awal dan akhir pengucapan.



Gambar 6 Keluaran sinyal dari fungsi penjendela jenis Hamming.



Gambar 7 Cepstrum untuk sinyal wicara "One"

Setelah fitur dari sinyal ini diperoleh dan dikuantisasi, maka fitur ini akan dimodelkan dengan pemodelan HMM dan disimpan ke dalam basis data. Sehingga untuk tiap-tiap sinyal wicara yang berbeda akan mempunyai model kata yang berbeda pula. Pada saat pengenalan, pengambilan keputusan dilakukan setelah ditemukan akhir dari pengucapan sinyal wicara. Untuk setiap berakhirnya suatu pengucapan kata, maka proses perbandingan kemiripan dilakukan antara kata yang diujikan dengan pustaka kata yang ada pada basis data. Untuk tiap-tiap model kata yang diuji, sistem akan melakukan penilaian pada semua model kata yang ada pada pustaka kata. Banyaknya *state* pada model HMM tergantung dari panjang durasi katanya. Setiap terjadi transisi pada tiap-tiap *state*, maka akan dilakukan penambahan nilai pada model kata tersebut. Proses transisi dilakukan dengan berdasarkan kemiripan karakteristik setiap *state*.

Pada fase pengujian tiap kata diuji sebanyak 35 kali. Hasil pengujian sistem dengan metode cepstrum dibandingkan dengan hasil pengujian sistem dengan metode spektrum. Hasil pengujian sistem pengenal wicara untuk setiap kata ditunjukkan pada tabel 1. Rata-rata hasil keberhasilan yang didapat sistem dengan metode cepstrum adalah 76,52% dengan standar deviasi 9,43%, sedangkan rata-rata hasil keberhasilan yang didapat sistem dengan metode spektral adalah 61,61% dengan standar deviasi 11,42%.

Tabel 1. Akurasi Sistem Pengenal Wicara

Kata	Akurasi dengan Metode Cepstrum	Akurasi dengan Metode Spektrum
"One"	82,86%	65,71%
"Two"	85,71%	71,42%
"Three"	82,86%	51,43%
"Four"	77,14%	77,14%
"Five"	77,14%	65,71%
"Six"	80,00%	80,00%
"Seven"	82,86%	51,43%
"Eight"	88,57%	82,86%
"Nine"	77,14%	74,29%
"Zero"	68,57%	48,57%
"Computer"	77,14%	60,00%
"Lamp"	68,57%	51,43%
"Fan"	65,71%	71,43%
"On"	85,71%	57,14%
"Off"	54,29%	48,57%
"Next"	71,43%	71,43%
"Previous"	88,57%	62,85%
"Shutdown"	85,71%	45,71%
"Hello"	82,86%	54,29%
"Stop"	80,00%	65,71%
"Music"	77,14%	51,43%
"Notepad"	68,57%	45,71%
"Email"	71,43%	62,86%
Rata-rata	**76,52%**	**61,61%**
Standar Deviasi	**9,43%**	**11,42%**

6. Kesimpulan

Dari sistem yang telah dirancang dan direalisasikan dapat disimpulkan bahwa sistem pengenal wicara dengan metode cepstral dan pemodelan HMM mampu mengenali 23 kata dengan tingkat keberhasilan rata-rata 76,52% dan standar deviasi 9,43%. Sistem pengenal wicara dengan metode cepstral mempunyai tingkat keberhasilan rata-rata yang lebih tinggi daripada sistem pengenal wicara dengan metode spektral.

Daftar Pustaka

1. T.W. Parsons *"Voice and Speech Processing"*, McGraw-Hill, 1976
2. A.V. Oppenheim, R.W. Schafer *"Discrete-Time Signal Processing"* Prentice Hall, 1999.

3. L.R. Rabiner, B.H. Juang, "*An Introduction to Hidden Markov Models*", *IEEE ASSP Magazine*, January 1986.
4. E. Karpov, "*Real Time Speaker Identification*", Department of Computer Science, University of Joensuu, Master Thesis, March 2003.
5. S. Furui, "*Digital Speech Processing, Synthesis and Recognition*", New York, Marcel [illegible]
6. L. R. Rabiner, "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*", *Proceeding of IEEE*, vol. 77, no. 2, pp. 257-286, Februari 1989.