Pengembangan Model Markov Tersembunyi untuk Pengenalan Kata Berbahasa Indonesia

Agus Buono[1], Yani Mandasari[2], Shelvie Nidya Neyman[3]

Departemen Ilmu Komputer FMIPA IPB
Kampus IPB Darmaga-Bogor
pudesha@yahoo.co.id

Abstrak

Pada paper ini disajikan suatu penerapan model HMM sebagai pengenal kata dengan ekstraksi ciri menggunakan teknik MFCC yang berbasis nilai power spektrum dari suara. Sistem yang dikembangkan bersifat text dependent dan melibatkan 10 pembicara yang mengucapkan 18 jenis kata. Pada penelitian, ada 3 jenis gugus data untuk melatih model HMM yang terdiri dari 4, 6 dan 8 hidden state, yaitu gugus yang terdiri suara laki-laki saja, gugus yang terdiri dari suara perempuan saja, dan gugus yang terdiri dari campuran suara laki-laki dan perempuan. Ada 4 jenis data uji, yaitu data uji suara laki-laki yang disertakan pada model pelatihan, data uji suara perempuan yang disertakan pada model, data uji suara laki-laki yang tidak disertakan pada model, dan data uji suara perempuan yang tidak disertakan pada model. Hasil percobaan menunjukkan bahwa sistem dapat mengenali kata dengan sangat baik (sekitar 98%), kalau diucapkan oleh pembicara yang disertakan dalam pembuatan model. Sistem gagal melakukan pengenalan untuk pembicara yang tidak disertakan dalam model pelatihan. Namun dengan memperluas data pelatihan, hasil pengenalan meningkat sekitar 30 % dari sebelumnya. Dari aspek jumlah hidden state, secara umum terlihat bahwa jumlah hidden 8 memberikan akurasi yang lebih baik disbanding 4 atau 6.

Kata Kunci : Hidden Markov Model (HMM), Me-Frequency Cepstrum Coefficients (MFCC), Sistem Pengenalan Kata (SPK).

I. Pendahuluan

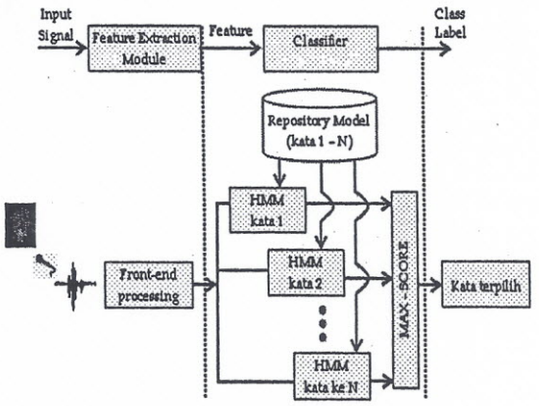
Sistem Pengenalan Kata (SPK), adalah suatu sistem pengenalan suara yang mengidentifikasi kata atau frase yang diucapkan oleh seorang pembicara. Dalam perkembangan metodologi, teknik pemodelan suara yang banyak dikaji adalah yang berbasis teori peluang. Satu teknik yang telah menunjukkan efektifitas yang baik dalam merepresentasikan suara adalah HMM (Hidden Markov Model), seperti disajikan pada [1].

Dari aspek ekstraksi ciri, *Mel-Frequency Cepstrum Coefficients* (*MFCC*) merupakan teknik yang telah luas dipakai pada pemrosesan sinyal suara, terutama pada pengenalan pembicara. Penggunaan teknik ini pada sistem pemrosesan sinyal memberikan pengenalan yang lebih baik dibandingkan dengan metode lainnya, Davis and Mermelstein (dalam [2]).

Paper ini disajikan dengan susunan sebagai berikut : Bagian 2 mengenai prinsip sistem identifikasi kata. Teknik analis fitur suara dan HMM disajikan pada bagian 3. Bagian 4 menyajikan data, rancangan dan hasil percobaan, dan sebagai penutup adalah kesimpulan dan saran untuk penelitian selanjutnya yang disajikan pada bagian 5.

2. Prinsip Sistem Pengenalan Kata

Secara umum, sistem pengenalan kata terdiri dari dua subsistem, yaitu subsistem ekstraksi ciri dan subsistem pencocokan pola, seperti disajikan pada **Gambar 1.** Subsistem ekstraksi ciri melakukan proses transformasi sinyal input ke dalam satu set vektor ciri sebagai representasi dari sinyal suara. Subsistem pencocokan pola merupakan bagian untuk melakukan identifikasi suara yang belum diketahui "kata apa yang diucapkan" dengan cara membandingkan sinyal suaranya yang telah diekstrak ke dalam vektor ciri dengan set vektor ciri dari "kata" yang telah diketahui dan tersimpan dalam sistem.



Gambar 1. Blok diagram sistem pengenalan kata dengan HMM sebagai pengenal pola

3. Analisis Fitur Suara dan HMM

Analisis Fitur Suara

Input dari analisis fitur suara adalah sinyal suara analog dan sebagai outputnya adalah *feature vector* untuk setiap *frame* (*time slice*). Tahap pertama adalah melakukan digitasi terhadap sinyal suara analog (disebut sebagai *analog-to-digital conversion*). Proses ini terdiri dari *sampling* dan kuantisasi, [3].

Sampling artinya mengukur amplitudo sinyal pada suatu indeks waktu tertentu. Dalam hal ini dikenal istilah *sampling rate*, yaitu banyaknya *sampling* yang dilakukan setiap detik. *Sampling rate* biasanya berkisar 8000 hingga 20000 *sample* per detik. Berikutnya adalah kuantisasi, yaitu menyimpan nilai amplitudo ke dalam nilai integer, yang dalam hal ini memakai representasi 8 bit atau 16 bit.

Setelah sinyal didigitasi, berikutnya adalah menyekatnya ke dalam *frame* dan menkonversikannya menjadi *feature vector* yang selanjutnya menjadi masukan bagi tahap berikutnya.

Fitur yang dipakai dalam penelitian ini adalah *Mel Frequency Cepstral Coeeficients* (MFCC).MFCC merupakan fitur yang populer saat ini. MFCC didasarkan pada variasi dari frekuensi kritis telinga manusia. Filter diletakkan secara linear pada frekuensi rendah dan logaritmik pada frekuensi tinggi untuk mendapatkan karakteristik suara yang penting. Diagram blok yang merepresentasikan struktur MFCC dapat dilihat pada Gambar 2, [4].

Dari Gambar 2 terlihat empat tahapan dalam ekstraksi ciri menggunakan MFCC, yaitu :

Frame blocking: sinyal suara dibaca per blok (frame) yang terdiri dari N sample. Antara dua frame yang bersisihan terdapat overlap N-M sample, dengan M adalah banyaknya pergeseran antar frame (M<N).

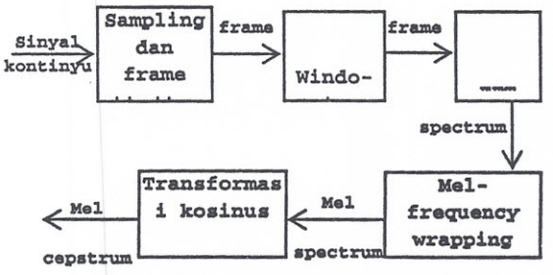
Windowing: proses windowing dilakukan pada setiap frame dengan tujuan untuk meminimumkan diskontinuitas antar sua frame, khususnya pada bagian awal dan akhir.

FFT (Fast-Fourier Transform): Pada tahap ini setiap frame yang terdiri dari N samples dikonversi dari domain waktu ke domain frekuensi. Output dari proses ini disebut dengan nama spektrum atau periodogram.

Mel-Frequency wrapping: tahap ini merupakan proses pengfilteran dari spektrum setiap frame yang diperoleh dari tahapan sebelumnya. Filter tersebut berupa M filter segitiga sama tinggi dengan tinggi satu. Filter ini dibuat dengan mengikuti persepsi telinga manusia dalam menerima suara. Persepsi ini dinyatakan dalam skala 'mel' (berasal dari Melody) yang mempunyai hubungan tidak linear dengan frekuensi suara, [4]. Dalam hal ini skala mel-frequency adalah linear untuk frekuensi kurang dari 1000 Hz dan logaritmik untuk frekuensi di atas 1000 Hz. Satu relasi antara frekuensi bunyi (dalam Hz) dengan skala mel adalah, [4], [5] :

$$\hat{f}_{mel} = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \ldots\ldots (1)$$

Penjelasan detail mengenai teknik MFCC dapat dijumpai pada [2] dan [4].



Gambar 2. Block diagram teknik MFFC

Windowing: proses windowing dilakukan pada setiap frame dengan tujuan untuk meminimumkan diskontinuitas antar sua frame, khususnya pada bagian awal dan akhir.

FFT: Pada tahap ini setiap frame yang terdiri dari N samples dikonversi dari domain waktu ke domain frekuensi. Output dari proses ini disebut dengan nama spektrum atau periodogram.

Mel-Frequency wrapping: tahap ini merupakan proses pengfilteran dari spektrum setiap frame yang diperoleh dari tahapan sebelumnya. Filter tersebut

berupa M filter segitiga sama tinggi dengan tinggi satu. Filter ini dibuat dengan mengikuti persepsi telinga manusia dalam menerima suara. Persepsi ini dinyatakan dalam skala 'mel' (berasal dari Melody) yang mempunyai hubungan tidak linear dengan frekuensi suara, [4]. Dalam hal ini skala mel-frequency adalah linear untuk frekuensi kurang dari 1000 Hz dan logaritmik untuk frekuensi di atas 1000 Hz. Satu relasi antara frekuensi bunyi (dalam Hz) dengan skala mel adalah, [4], [5] :

$$\hat{f}_{mel} = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad \text{.....} \quad (1)$$

Cepstrum: Pada tahap ini dilakukan konversi dari koefisien spektrum mel kembali ke domain waktu menggunakan transformasi kosinus sesuai rumus 3.

$$C_j = \sum_{i=1}^{M} X_i * \cos\left(\frac{j*(i-0.5)*\pi}{20}\right) \quad \text{.....} \quad (2)$$

Dengan j=1,2,3,...,K; K adalah banyaknya koefisien; M adalah banyaknya filter segitiga; X_i adalah koefisien spektrum mel yang diperoleh dengan formula (1). Dalam hal ini C_j disebut sebagai mel frequency cepstrum coefficients (MFCC) koefisien ke j.

Hidden Markov Model

Hidden Markov Model (HMM), atau model Markov tersembunyi, ialah suatu model peluang temporal yang menggambarkan keterkaitan antar peubah *state* (*state variable*) dari waktu ke waktu, serta antara peubah *state* dengan peubah teramati (*observable variable*). Secara visual, model ini dapat digambarkan menggunakan suatu *finite state automata* dengan banyaknya *state* adalah sesuai dengan banyaknya kemungkinan kombinasi nilai variabel dalam model. Dalam hal ini, setiap *state* merupakan suatu kombinasi variabel tesebut. Sebagai contoh, jika terdapat suatu model temporal dengan tiga variabel biner maka banyaknya *state* adalah $2^3 = 8$ buah. Di dalam HMM, peubah *state* adalah peubah yang tak teramati (*hidden variable*), dan peubah yang teramati (*observable variable*).

Berikut adalah notasi yang digunakan dalam HMM, [5] :

N : Banyaknya *hidden state* (*state* ke 1, 2, 3, ..., n). Sedangkan q_t menotasikan *state* ke-q pada indeks waktu t.

M : Banyaknya kemungkinan kemunculan peubah teramati. Sedangkan v_k, untuk k=1, 2, 3, ..., M, adalah nilai-nilai peubah teramati.

Π : adalah $\{\pi_i\}$, dengan $\pi_i = P(q_1 = i)$, yaitu peluang pada tahap awal berada pada *statei*. Dalam hal ini $\sum_{i=1}^{N} \pi_i = 1$

A : adalah $\{a_{ij}\}$ dengan $a_{ij} = P(q_{t+1} = j | q_t = i)$, yaitu peluang berada di *state* j pada waktu t+1 jika pada waktu t berada di state i. Dalam hal ini diasumsikan a_{ij} bebas dari waktu.

B : adalah $\{b_j(k)\}$, dengan $b_j(k) = P(v_k$ pada waktu $t | q_t = j)$, yaitu peluang peubah teramati yang muncul adalah simbol v_k.

O_t : adalah notasi untuk nilai teramati pada waktu t, sehingga barisan nilai teramati (*observable symbol*) adalah $O = O_1, O_2, O_3, \ldots, O_T$. Dengan T adalah panjang observasi yang dilakukan.

Dengan notasi-notasi seperti di atas, maka suatu HMM dilambangkan dengan :

$$\lambda = (A, B, \Pi)$$

Secara umum ada tiga masalah dasar yang terdapat dalam HMM, [5], yaitu : (1) Evaluasi untuk menduga peluang munculnya barisan $O = O_1, O_2, O_3, \ldots, O_T$ dari sebuah HMM; (2) Decoding untuk memilih barisan state $Q = q_1, q_2, \ldots, q_T$ yang 'optimal', yaitu yang paling besar kemungkinannya menghasilkan O yang diketahui; dan (3) Pembelajaran parameter HMM, yaitu melakukan pendugaan terhadap parameter-parameter model HMM, $\lambda = (A,B,\Pi)$, sehingga $P(O|\lambda)$ atau $P(O,Q|\lambda)$ maksimum. Secara detail, ketiga algoritma tersebut dapat dijumpai di [1] dan [5].

4. Rancangan Percobaan dan hasil

Rancangan Percobaan

Data yang digunakan adalah gelombang suara yang direkam dari 10 pembicara, yaitu 5 laki-laki (pembicara1, 2, 3, 7, dan 8) dan 5 perempuan(pembicara 4, 5, 6, 9, dan 10) dengan rentang umur20-24 tahun. Data tersebut disimpan dalam *file* berekstensi WAV.

Data pelatihan diperoleh dari pembicara 1-6 yang diminta untuk mengucapkan 18 kata. Sistem yang dikembangkan untuk mengenali kata-kata tertentu seprti disajikan pada Tabel 1.

Tabel 1 Daftar kata-kata yang digunakan dalam penelitian.

Kelompok Fonem	Posisi Fonem		
	Awal	Tengah	Akhir
/i/	Ikan	Pintu	Padi
/e/	Ekor	Nenek	Sore
/ə/	Emas	Ruwet	Tante
/a/	Anak	Kantor	Kota
/u/	Ukir	Tunda	Baru
/o/	Obat	Kontan	Baso

Data pengujian dibagi menjadi 4 kelompok: data tes 1, data tes 2, data tes 3, dan data tes 4. Pembagian ini berdasarkan pada perbedaan jenis kelamin dan keikutsertaan pembicara dalam pelatihan. Data tes 1 dan data tes 2 berasal dari speaker 1-6 dengan 3 kali pengulangan untuk setiap kata.

Data tes 3 dan data tes 4 berasal dari pebmicara 7, 8, 9 dan 10 dengan 5 kali pengulangan untuk setiap kata. Tabel 2 menyajikan proporsi pembagian data untuk pelatihan dan pengujian.

Tabel 2 Proporsi pembagian data untuk pelatihan dan pengujian.

Speaker	Jumlah File Pelatihan	Jumlah File Pengujian	Kelompok Pengujian
1	7	3	Data tes 1
2	7	3	
3	7	3	
4	7	3	Data tes 2
5	7	3	
6	7	3	
7	-	5	Data tes 3
8	-	5	
9	-	5	Data tes 4
10	-	5	

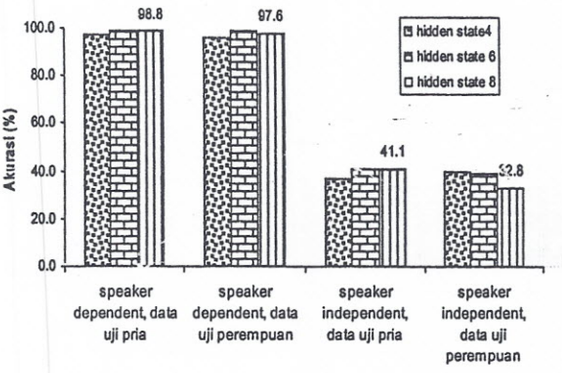
Analisis fitur suara MFCC (*Mel-Frequency Cepstral Coefficients*) diimplementasikan dengan menggunakan *Auditory Toolbox* yang dikembangkan oleh Slanley pada tahun 1998. *Auditory Toolbox* dapat diperoleh secara bebas dihttp://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/), [6]. Data suara dalam percobaan ini merupakan data *mono* (satu saluran) yang didigitasi dengan *bit rate* sebesar 16-bit dan *sampling rate* 16000 Hz, karena pada umumnya *sampling rate* yang digunakan oleh mikrofon *wideband* berada pada 16000 Hz. Langkah selanjutnya adalah membagi gelombang suara ke dalam *frame* dengan 100 sampel tiap *frame*-nya, hal ini sesuai dengan standar yang terdapat dalam *Auditory Toolbox*. Melalui proses MFCC, maka akan dihasilkan 13 koefisien *mel cepstrum* untuk tiap *frame*.

Pada penelitian ini, jenis HMM yang digunakan adalah HMM *left-right*, dengan jumlah hidden state yang dicobakan adalah 4, 6 dan 8 . Parameter HMM diduga dengan algoritma *Segmental K-means*, yang secara detail disajikan pada [1] dan [5]. Selain jumlah hidden state, ada 3 jenis data pelatihan, yaitu

- Data pelatihan 1, Model HMM dilatih dengan data latih dari pembicara laki-laki saja.
- Data pelatihan 2, Model HMM dilatih dengan data latih dari pembicara perempuan saja.
- Data pelatihan 3, Model kata dilatih dengan campuran suara laki-laki dan perempuan.

Hasil dan Pembahasan

Gambar 3 menyajikan perbandingan hasil akurasi dari perbagai kondisi data latih dan data uji untuk model HMM dengan jumlah hidden state sebanyak 4, 6 dan 8. Grafik paling kiri adalah untuk data latih laki-laki dan diuji dengan data uji laki-laki dari orang yang suaranya dipergunakan untuk pelatihan model. Posisi ke dua adalah kondisi yang sama dengan sebelumnya, hanya saja jenis kelamin pembicaranya adalah perempuan. Dari sini terlihat bahwa untuk kedua kondisi tersebut, yaitu speaker dependent, sistem dapat melakukan pengenalan dengan baik, yaitu rata-rata sekitar 97.5%.

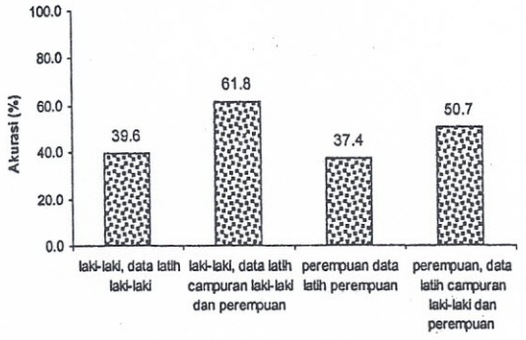


Gambar 3. Perbandingan akurasi sistem untuk berbagai kondisi data latih dan data uji

Posisi ke tiga dan ke empat adalah untuk pembicara laki-laki dan perempuan, namun suara yang diuji bukan dari pembicara yang disertakan pada pelatihan. Hasil percobaan menunjukkan bahwa sistem gagal melakukan pengenalan dengan baik, dengan akurasi sekitar 40% untuk laki-laki dan 32% untuk perempuan. Fakta ini menunjukkan bahwa sistem yang dibangun masih bersifat speaker dependent, dan gagal untuk kondisi speaker independent. Untuk kasus speaker dependent, terlihat bahwa jenis kelamin tidak memberikan pengaruh terhadap hasil akurasi. Dalam hal ini kedua kondisi tersebut memberikan akurasi yang tinggi (>95%). Sedangkan untuk kasus speaker independent, meskipun secara akurasi masih rendah, namun terlihat bahwa suara laki-laki lebih mudah dikenali. Hal ini menunjukkan bahwa variasi antar suara laki-laki tidak terlalu besar dibandingkan dengan suara dari perempuan.

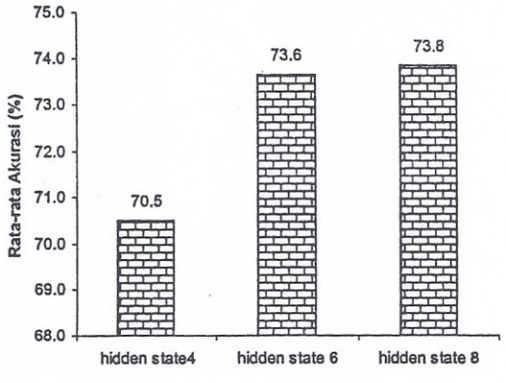
Salah satu pendekatan yang dilakukan untuk mengatasi masalah tersebut adalah dengan menambah jumlah pembicara yang disertakan dalam pelatihan. Gambar 4 menyajikan

perbandingan akurasi antara model dengan data latih terbatas dan dan model dengan data latih diperbesar cakupannya.



Gambar 4. Perbandingan akurasi sistem untuk kondisi independent speaker untuk berbagai kondisi data latih dan data uji

Dari Gambar 4 terlihat bahwa dengan penambahan pembicara sebagai data latih, akurasi sistem meningkat hampir 20% untuk data uji laki-laki dan sekitar 13% untuk data uji perempuan. Dari fakta ini ada 2 hal yang bisa disebutkan, yaitu bahwa penambahan pembicara yang disertakan pada pelatihan akan meningkatkan akurasi sistem yang bersifat independent speaker. Kedua adalah memperkuat pernyataan sebelumnya yang menyatakan bahwa suara laki-laki lebih mudah dikenali dibanding suara perempuan.



Gambar 5. Perbandingan rata-rata akurasi dari semua kondisi untuk berbagai jumlah hidden state HMM

Dari segi jumlah hidden state pada model HMM, terlihat bahwa HMM dengan hidden state sebanyak 8 memberikan akurasi terbaik, yang secara rata-rata dari semua jenis percobaan memberikan akurasi sebesar 73.8%. Nilai ini sedikit di atas HMM dengan jumlah hidden state sebanyak 6. Untuk HMM dengan jumlah hidden state 4, terlihat bahwa sistem kurang mampu melakukan pengenalan dengan baik, yaitu dengan rata-rata akurasi 70.5%.

5. Kesimpulan

Beberapa hal yang dapat disimpulkan dari penelitian ini adalah :

1. Model MFCC sebagai ekstraksi ciri dan HMM sebagai pengenal pola mampu diterapkan pada sistem identifikasi kata yang bersifat speaker dependent dengan akurasi berkisar 97.5%.
2. Peningkatan akurasi untuk kondisi independent speaker dapat dilakukan dengan menambah pembicara yang disertakan dalam model. Hasil percobaan menunjukkan peningkatan yang cukup berarti, yaitu sekitar 20% untuk pembicara laki-laki dan 13% untuk pembicara perempuan.
3. Secara umum dapat disimpulkan bahwa suara laki-laki relatih lebih mudah dikenali dibanding dengan suara perempuan.
4. Jumlah hidden state HMM yang layak pada sistem pengenalan kata adalah sebanyak 8 buah.

6.Referensi

[1] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceeding IEEE, Vol 77 No. 2, pp 257-289, 1989.

[2] Todor D. Ganchev. *Speaker Recognition*. PhD Dissertation, Wire Communications Laboratory, Department of Computer and Electrical Engineering, University of Patras Greece. 2005.

[3] Jurafsky D, Martin JH. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice Hall.

[4] Cornaz, C. dan U. Hunkeler. An Automatic Speaker Recognition System. Mini-Project. *http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition*, access : August, 15, 2005.

[5] Dugad R, Desai UB. 1996. *A Tutorial on Hidden Markov Models*. Technical Report, Department of Electrical Engineering, Indian Institute of Technology – Bombay, India.

[6] Do MN. 1994. *Digital Signal Processing Mini-Project: An Automatic Speaker Recognition System*. Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland. http://lcavwww.epfl.ch/~minhdo/asr_project/asr_project.pdf [27 September 2005]