APLIKASI *SPEECH TO TEXT* BERBAHASA INDONESIA MENGGUNAKAN *MEL FREQUENCY CEPSTRAL COEFFICIENTS* DAN *HIDDEN MARKOV MODEL* (HMM)

Eko Widiyanto, Sukmawati Nur Endah, Satriyo Adhy, Sutikno
Jurusan Ilmu Komputer/Informatika, FSM, Universitas Diponegoro

ABSTRAK

Dalam menyampaikan informasi. Informasi dapat disampaikan dalam bentuk teks, citra, suara, dan lain-lain. Pemakaian teknologi komputer untuk menjalin komunikasi manusia dengan komputer sudah menjadi suatu kebutuhan, jika komputer mengerti ucapan yang dimaksudkan manusia akan menjadi suatu kemudahan dalam pengoprasian komputer. Pada penelitian *speech recognition* ini terdapat dua proses yang sangat penting yaitu *feature extraction* dan *template matching*. Dalam metode *feature extraction*-nya disini menggunakan *Mel Frequency Cepstral Coefficients* (MFCC) dan *Hidden Markov Model* (HMM) sebagai metode *template matching*-nya. Pengujian ini dilakukan dengan menggunakan nilai koefisien 8 pada MFCC dan 14 state pada HMM. Terdapat 21 kata yang akan dikenali dan 36 suku kata pembangunnya. Dataset yang dipakai berjumlah 1170 suku kata dan 210 kata dengan menggunakan 10 *speaker* yang berbeda. Berdasarkan parameter-parameter itu didapatkan akurasi terbaik 72.61 % berdasarkan hasil pengujian yang telah dilakukan. Komunikasi merupakan salah satu cara yang paling efektif untuk menyampaikan maksud dan tujuan

Kata kunci : Speech to text, MFCC, HMM

1. PENDAHULUAN

Komunikasi bahasa antar manusia dengan manusia merupakan salah satu cara yang paling efektif untuk menyampaikan maksud dan tujuan seseorang dalam menyampaikan informasi untuk memudahkan seseorang dalam menyelesaikan pekerjaan. Informasi dapat disampaikan dalam bentuk teks, citra, suara, dan lain-lain. Pemakaian teknologi komputer untuk menjalin komunikasi manusia dengan komputer sudah menjadi suatu kebutuhan, jika komputer mengerti ucapan yang dimaksudkan manusia akan menjadi suatu kemudahan dalam pengoprasian komputer, seperti *voice command*, akses kontrol sistem berbasis suara, dan identifikasi suara untuk keamanan sistem.

Perkembangan *speech recognition (speech to text)* berjalan cukup pesat pada saat ini dilihat dari banyaknya jurnal yang membahas mengenai *speech to text* untuk bahasa Inggris. Suara manusia mempunyai karakteristik yang sangat kompleks, satu kata yang diucapkan oleh orang yang berbeda akan menghasilkan karakteristik suara yang berbeda, namun suatu sistem diharuskan dapat mengenali sebagai suatu kata yang sama. Selain itu faktor yang mempengaruhi suara adalah kesehatan, psikologi, umur, dan jenis kelamin seseorang.

speech to text memungkinkan suatu perangkat untuk mengenali dan memahami kata-kata yang diucapkan dengan cara digitalisasi kata dan mencocokkan sinyal digital tersebut dengan suatu pola tertentu yang tersimpan dalam suatu perangkat. Kata-kata yang diucapkan diubah bentuknya menjadi sinyal digital dengan cara mengubah gelombang suara menjadi sekumpulan angka yang kemudian disesuaikan dengan kode-kode tertentu untuk mengidentifikasikan kata-kata tersebut. Hasil dari identifikasi kata yang diucapkan dapat ditampilkan dalam bentuk tulisan.

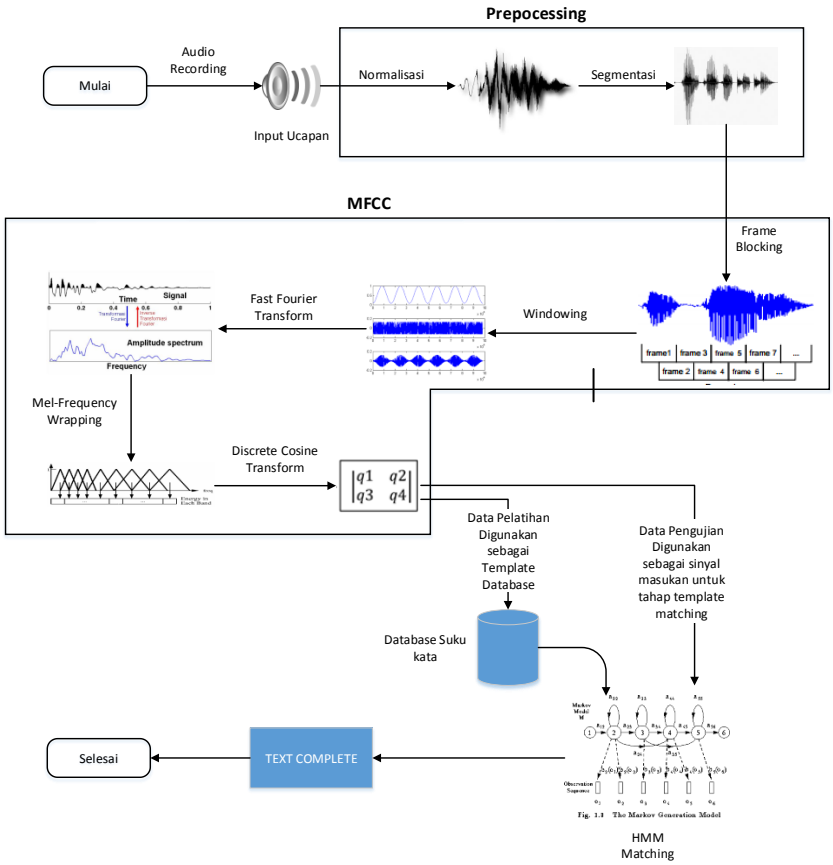
Di Indonesia penelitian mengenai *speech to text* sudah mulai banyak dilakukan dilihat dari bermunculannya jurnal-jurnal mengenai *speech to text* bahasa Indonesia. Dari penelitian sebelumnya ada yang menggunakan metode ekstrasi ciri *Mel Frequency Cepstral Coefficients* (MFCC) dan metode pengenalan pola *hidden markov model* (HMM) [Fawziah, 2013], namun masih terbatas untuk speaker laki – laki saja dan hanya beberapa kata saja yang menjadi data pengujian.Dalam penelitian kali ini dicoba untuk menggunakan metode yang sama dengan menggunakan speaker laki-laki dan perempuan dengan data pengujian yang lebih banyak.

Mel Frequency Cepstrum Coefficients (MFCC) feature extraction mengkonversikan sinyal suara kedalam beberapa vektor data berguna bagi proses pengenalan suara. Terdapat 7 tahapan dalam MFCC yaitu *Pre Emphasize, Frame Blocking, Windowing, Fast Fourier Transform, Mel Frequency Warping, Discrete Cosine Transform, dan Cepstral Liftering*. Metode ini memiliki beberapa kelebihan diantaranya, menangkap informasi penting dalam sinyal suara, menghasilkan data seminimal mungkin tanpa menghilangkan informasi, dan mereplikasikan organ pendengaran manusia dalam melakukan persepsi terhadap sinyal suara [Andriana, 2011].

Hidden Markov Model (HMM) merupakan suatu metode pendekatan yang dapat mengelompokan sifat-sifat spectral dari tiap bagian suara dengan beberapa pola. HMM memiliki 5 proses dasar dalam melakukan pengenalan suara, yaitu: *Feature Analysis, Unit Matching System, Lexical Decoding, Systactic Analysis, Semantic Analysis*. Proses-proses itulah yang menyebabkan HMM mempunyai tingkat akurasi

yang lebih tinggi dibanding metode lain, terbukti dengan banyaknya penelitian mengenai *speech recognition* [Rabiner, 1993].

Oleh karena itu dalam penelitian ini akan dilakukan pemodelan aplikasi *speech to text* berbahasa Indonesia menggunakan metode ekstrasi ciri *Mel-Frequency Cepstral Coefficient* (MFCC) dengan menggunakan *Hidden Markov Model* (HMM) untuk mengenali pola ucapannya.



Gambar 1. Arsitektur aplikasi *speech to text* berbahasa Indonesia

2. ARSITEKTUR SISTEM

Arsitektur sistem yang dibuat yaitu seperti pada gambar 1.

Ekstraksi ciri pada arsitektur sistem pada gambar 1 diatas dengan menggunakan metode *Mel Frequency* Cepstral *Coefficients* sedangkan pengenalan suara, *parsing/chunking*, ekstraksi informasi, dan peringkasan teks dengan menggunakan *Hidden markov model* (HMM).

Mel Frequency Cepstral Coefficients

Mel-Frequency *Cepstral Coefficient* (MFCC) merupakan metode eksrasi ciri yang menghitung koefisien *cepstral* dengan mempertimbangkan pendengaran manusia.

Keunggulan dari metode ini adalah [Putra at al, 2011]:

1. Mampu untuk menangkap karakteristik suara yang sangat penting bagi pengenalan suara, atau dengan kata lain, mampu menangkap informasi-informasi penting yang terkandung dalam sinyal suara.
2. Menghasilkan data seminimal mungkin tanpa menghilangkan informasi-informasi penting yang ada.
3. Mengadaptasi organ pendengaran manusia dalam melakukan persepsi terhadap sinyal suara.

Proses MFCC mempunyai beberapa tahapan, yaitu:

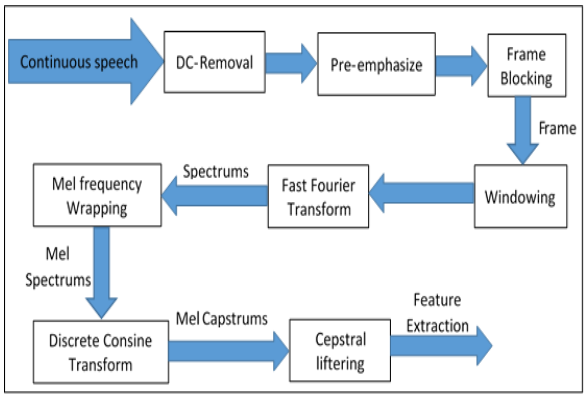
1. *Preprocessing*

   Preprocessing pada MFCC meliputi *DC-removal, Pre-emphasize, Frame Blocking*, dan *Windowing*.

   DC-Removal berfungsi untuk mendapatkan nilai normalisasi dari data suara input, dengan menghitung rata-rata dari data sampel suara. *Preemphasize* bertujuan untuk mengurangi *noise ratio* pada sinyal dan menyeimbangkan spectrum dari *voice sound*. Proses *Frame Blocking* berfungsi untuk memotong sinyal suara dengan durasi yang panjang menjadi durasi yang lebih pendek, agar mendapatkan karakeristik sinyal yang *periodic*. Proses *windowing* bertujuan untuk

mengurangi kebocoran spectral atau aliasing yang merupakan efek dari *frame blocking* yang menyebabkan sinyal menjadi *discontinue*.

2. FFT (Fast Fourier Transform)
   FFT merupakan metode transformasi untuk mendapatkan sinyal dalam domain frekuensi dari sinyal diskrit yang ada.
3. *Mel-Frequency Wrapping*
   Filterbank dilakukan dengan tujuan untuk mengetahui energi dalam sinyal suara. Frekuensi dalam sebuah sinyal diukur menggunakan mel scale.
4. *Cepstrum*
   Mel-Frequency Cepstrum didapatkan dari proses DCT untuk mendapatkan kembali sinyal dalam domain waktu. Hasilnya disebut sebagai Mel-Frequency Cepstral Coefficient (MFCC).
5. *Cepstral Filtering*
   Hasil dai MFCC mempunyai beberapa kelemahan yaitu pada low-order yang sangat sensitive terhadap spectral slope dan high-order yang sangat sensitivv terhadap noise. Oleh karena itu, maka cepstral filtering menjadi salah satu metode untuk meminimalisasi sensitifitas tersebut.

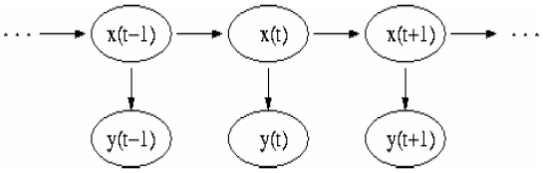


Gambar 2. Block Diagram Untuk MFCC

Hidden Markov Model (HMM)

Hidden markov model (HMM) merupakan pendekatan yang dapat mengelompokan sifat-sifat spectral dari tiap bagian suara pada beberapa pola. Teori dasar dari HMM adalah dengan mengelompokan sinyal suara sebagai proses parametric acak, dan parameter proses tersebut dapat dikenali (diperkirakan) dalam akurasi yang tepat.

Diagram pada gambar 3 menunjukkan arsitektur umum dari HMM, seperti yang disajikan pada gambar 3. Tiap bentuk oval memiliki variabel random yang dapat mengambil nilai. Variabel random x(t) yaitu nilai dari variabel tersembunyi pada waktu t. Variabel random y(t) yaitu nilai variabel yang diteliti pada waktu t. Tanda panah pada diagram menunjukan ketergantungan kondisi. Dari diagram, ini jelas bahwa nilai variabel tersembunyi x(t) (pada waktu t) hanya tergantung pada nilai variable tersembunyi x(t-1) (pada waktu t-1). Serupa, nilai variabel yang diteliti y(t) hanya tergantung pada nilai variabel tersembunyi x(t) (keduanya pada waktu t).



Gambar 3. Evolusi temporal dari *Hidden Markov Model*

HMM memiliki lima komponen, yaitu:

1. Jumlah *state* (N)
   State merupakan parameter tersembunyi (*hidden state*). Pada penerapannya, jumlah *state* ini menjadi salah satu parameter uji. Jadi, jumlah *state* diset sedemikian rupa hingga didapatkan keluaran yang optimal. Karena model yang digunakan adalah HMM diskrit, maka *state* dilabelkan {1, 2, …, N} dan *state* pada waktu *t* dinotasikan sebagai *qt*.

2. Parameter model (M)
   Parameter model inilah yang merupakan parameter observasi (*observed state*). Pada aplikasi *speech recognition*, parameter model ini direpresentasikan oleh vektor ciri sinyal suara.

3. *Intial state* atau *state* awal ($\pi = \pi i$)

   $$\pi_i = P(q_t = i) \qquad 1 \le i \le N$$

4. Probabilitas transisi (A=[*aij*]), yaitu probabilitas dari perpindahan *state* i ke *state* j, dimana transisi antar *state*-nya dilakukan berdasarkan masukan observasi.

   $$a_{ij} = P(q_{t+1} = J | q_t = j) \; , \; 1 \le i, j \le N$$

5. Probabilitas simbol observasi ($B=\{b_j(k)\}$), yaitu probabilitas observasi yang dibangkitkan oleh *state*.

   $$b_j(k) = P(o_t = v_k | q_t = j), \; 1 \le j \le N, \; 1 \le k \le M$$

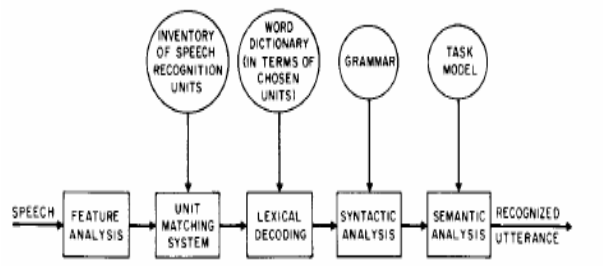
Berdasarkan kelima komponen di atas, untuk merancang HMM, dibutuhkan dua parameter model, yaitu N dan M. Selain itu, dibutuhkan tiga nilai probabilitas (π, A, B) yang dimodelkan dengan menggunakan notasi λ [$\lambda = (A, B \; \pi)$].

Menurut Rabiner, salah satu masalah yang dapat diselesaikan oleh HMM adalah masalah optimasi urutan *hidden state* berdasarkan urutan kejadian yang dapat diamati. Urutan *hidden state* optimal adalah urutan yang paling tepat yang "menjelaskan" kejadian

yang dapat diamati. Masalah ini disebut juga masalah *decoding*.

Implementasi HMM pada Pengenalan Suara

Salah satu implementasi HMM yang dijelaskan adalah pada sistem pengenalan suara. Diagram blok disajikan pada gambar 4. Gambar tersebut menunjukan diagram blok dari pendekatan pengenalan pola pada sistem pengenalan suara kontinyu.



Gambar 4. Diagram blok pengenalan suara kontinyu

Suara yang menjadi input pada gambar 4 akan melalui proses Feature Analysis yang memfilter suara input menjadi spekral-spektral suara. Setelah melalui proses Feature analysis, spectral suara kemudian akan dipecah menjadi suku kata – suku kata pada proses *Unit Matching System*. Pada proses *Unit Matching System*, sistem akan membaca database suku kata untuk kemudian dicari suku kata – suku kata yang mirip dengan spectral suara input. Pada *Lexical Decoding*, tiap suku kata yang terdapat di *Unit Matching System* disusun menjadi kata berdasarkan *Word Dictionary*. Pada *Synctactic Analysis*, tiap kata yang terdapat di *Lexical Decoding* disusun menjadi frase berdasarkan database frase grammar. Dengan berdasarkan pada *database Task Model*, *Semantic Analysis* memungkinkan pembentukan Kalimat dari frase-frase yang ada di *Systactic Analysis*.

3. PENGUJIAN

Sebelum dilakukan pengujian sistem di training terlebih dahulu. Training dilakukan pada pembicara 10 orang dengan 5 wanita dan 5 pria, dan 3 sample kata untuk setiap orang. Data uji dan training menggunakan 21 sample kata (“ada”, “ahmad”, “air”, “ayah”, “botol”, “cendol”, “dalam”, “di”, “dunia”, “ikan”, “jual”, “ke”, “kita”, “lenyap”, “mimpi”, “minum”, “om”, “pasar”, “pergi”, “telah” , “toko”).

Pengujian dilakukan untuk mengetahui akurasi sistem dengan menggunakan koefisien 8 pada MFCC dan state 14 pada HMM.

Tabel 1. Hasil Pengujian dengan data training suara laki-laki.

Data Uji	Laki-laki		Perempuan	
	A	B	A	B
Ada	√	√	-	-
Ahmad	√	√	-	-
Air	√	√	-	√
Ayah	√	√	√	-
Botol	√	√	√	-
Cendol	√	√	-	-
Dalam	√	√	-	-
Di	-	-	-	-
Dunia	√	√	-	-
Ikan	√	√	-	-
Jual	√	√	-	-
Ke	-	-	-	-
Kita	√	√	√	√
Lenyap	√	√	-	-
Mimpi	-	√	-	-
Minum	√	√	-	-
Om	-	-	-	-
Pasar	√	√	-	√
Pergi	√	√	√	√
Telah	√	√	-	-
Toko	√	√	√	-
∑error	4	3	16	17
Akurasi (%)	80.95	85.71	23.8	19.05
Rata-rata akurasi	83.33 %		21.42 %	

Tabel 2. Hasil Pengujian dengan data training suara perempuan.

Data Uji	Laki-laki		Perempuan	
	A	B	A	B
Ada	-	-	√	√
Ahmad	√	-	√	√
Air	-	-	√	-
Ayah	-	-	√	√
Botol	√	-	√	√
Cendol	√	√	√	√
Dalam	-	-	√	√
Di	-	-	-	-
Dunia	√	-	√	√
Ikan	-	√	√	√
Jual	-	-	√	√
Ke	-	-	-	-
Kita	-	√	√	√
Lenyap	-	√	√	√
Mimpi	-	-	-	-
Minum	-	-	√	-
Om	-	-	-	-
Pasar	√	-	√	√
Pergi	-	-	√	√
Telah	-	-	√	√
Toko	-	-	√	√
∑error	16	17	4	6
Akurasi (%)	23.8	19.05	80.95	71.42
Rata-rata akurasi	21.45 %		76.18 %	

Tabel 3. Hasil Pengujian dengan data training suara laki-laki dan perempuan.

Data Uji	Laki-laki		Perempuan	
	A	B	A	B
Ada	√	√	√	-
Ahmad	√	√	√	√
Air	√	√	√	√
Ayah	√	√	√	√
Botol	√	√	√	√
Cendol	√	√	√	√
Dalam	-	√	-	√
Di	-	-	-	-
Dunia	√	-	√	√
Ikan	√	√	-	√
Jual	√	√	√	√
Ke	-	-	-	-
Kita	√	√	√	√
Lenyap	√	√	-	√
Mimpi	√	√	-	√
Minum	-	√	√	√
Om	-	-	-	-
Pasar	√	√	√	√
Pergi	√	√	√	-
Telah	√	-	√	√
Toko	√	√	-	√
∑error	5	5	8	5
Akurasi (%)	76.19	76.19	61.90	76.19
Rata-rata akurasi	76.19 %		69.04 %	

4. ANALISA HASIL

Pada proses untuk pelatihan pembuatan database sistem dibagi menjadi 3 yaitu, dengan hanya menggunakan 5 speaker laki-laki, 5 speaker perempuan, dan menggunakan masing-masing 5 speaker untuk laki-laki dan perempuan. Hal ini dilakukan untuk mengetahui apakah jenis kelamin seseorang sangat mempengaruhi pendeteksian suara tau tidak. Berdasarkan hasil pengujian pendeteksian pada suara perempuan lebih sukar di kenali.

Tidak adanya standarisasi sinyal suara yang dilakukan pada penelitian ini membuat tingkat akurasi sistem menjadi turun jika diuji menggunakan speaker yang mempunyai suara sangat berbeda dengan suara yang dijadikan pelatihan.

Pada proses pengambilan sinyal suara perlu adanya standarisasi cara bicara agar sinyal suara yang dihasilkan mempunyai kemiripan yang lebih. Standarisasi gaya bicara diproses pemgambilan data juga bisa dijadikan sebagai standart pengujian yang lakukan. Agar sistem bias menghasilkan tingkat akurasi yang lebih akurat.

5. KESIMPULAN

Berdasarkan hasil analisis penelitian yang dilakukan, maka dapat ditarik kesimpulan akurasi maksimum dari sistem speech *to text* yang dirancang menggunakan ekstrasi ciri MFCC dan metode pencocokan HMM adalah 72.61 % dengan dipengaruhi oleh jumlah koefisien MFCC dan state pada HMM.

6. DAFTAR PUSTAKA

[1] Andriana A. D., 2011. "*Perangkat Lunak Untuk Membuka Aplikasi Pada Komputer Dengan Perintah Suara Menggunakan Metode Mel Frequency Cepstrum Coefficientss (MFCC)".* Tugas Akhir Universitas Komputer Indonesia. Bandung

[2] Arlow J. dan Neustad I. 2002. "*UML and The Unified Process Practical Object-Oriented Analysist & Design"*. Pearson Education Limited. Great Britain.

[3] Basuki T. A., 2000. *"Pengenalan Suku Kata Bahasa Indonesia Menggunakan Finite-State Automata"*. Universitas Katolik Parahyangan. Bandung

[4] Cole A., Mariani J., Uszkoreit H., et al. eds*."Survey of the state of the art in human language technology"*. Cambridge Studies In Natural Language Processing. XII–XIII. Cambridge: University Press. ISBN 0-521-59277-1.

[5] Departemen Pendidikan Nasional., 2009. *"Pedoman umum ejaan bahasa Indonesia yang disempurnakan"*. Jakarta: Pusat Bahasa

[6] Faisal M., 2009. *"Kajian Bahasa Indonesia SD: Struktur Fonologi Dan Morfologi Bahasa Indonesia"*. Jakarta: Direktorat Jendral Pendidikan Tinggi Departemen Pendidikan Nasional

[7] Fawziah S.K., 2013*."Pemodelan Speech Recognition Speech-To-Text Dalam Bahasa Indonesia Menggunakan Mel Frequency Cepstral Coefficients (MFCC) Dan Hidden Markov Model (HMM)"*. Tugas Akhir Institut Teknologi Telkom. Bandung.

[8] Firdian F., 2011. *"Pengontrolan Elevator Berbasis Sistem Pengenalan Ucapan"*. Tugas Akhir Universitas Komputer Indonesia. Bandung.

[9] Melissa, G., 2008. *"Pencocokan Pola Suara (Speech Recognition) dengan Algoritma FFT Dan Divide And Conquer"*. Tugas Akhir Institut Teknologi Bandung. Bandung

[10] Mulya A., Utama A.A., dan Putra A.M., 2007. *"Analisa dan Perancangan Perangkat Lunak Perintah Suara Sebagai Penunjang Sarana Input Pada Sistim Operasi Microsoft Windows XP"*. Tugas Akhir Universitas Bina Nusantara. Jakarta.

[11] Mutiara, D., 2009. "Penentuan pitch sinyal ucapan huruf vocal pria dan wanita dalam bahasa Indonesia". Teknik Elektro Universitas Semarang. Semarang

[12] Putra, D. dan Resmawan, A. 2011. "*Verifikasi Biometrika Suara Menggunakan Metode MFCC dan DTW".* Tugas Akhir Universitas Udayana. Bali.

[13] Rabiner, L., 1989. "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"*. Prentice-Hall, New Jersey.

[14] Rabiner L., and Juang B.H., 1993. *"Fundamentals of Speech Recognition"*. Prentice Hall International, Inc.

[15] Wibisono, Y., 2008. *"Penggunaan Hidden Markov Model untuk Kompresi Kalimat"*. Tesis Institut Teknologi Bandung. Bandung.