



**UNIVERSITAS INDONESIA**

**PEROLEHAN INFORMASI DOKUMEN SUARA  
PEMBICARAAN BERDASARKAN HASIL DARI SISTEM  
PENGENALAN SUARA UNTUK BAHASA INDONESIA**

**SKRIPSI**

**Armando Yonathan**

**1205000177**

**FAKULTAS ILMU KOMPUTER**

**PROGRAM ILMU KOMPUTER**

**DEPOK**

**JULI, 2009**



**UNIVERSITAS INDONESIA**

**PEROLEHAN INFORMASI DOKUMEN SUARA  
PEMBICARAAN BERDASARKAN HASIL DARI SISTEM  
PENGENALAN SUARA UNTUK BAHASA INDONESIA**

**SKRIPSI**

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer

**Armando Yonathan**

**1205000177**

**FAKULTAS ILMU KOMPUTER**

**PROGRAM ILMU KOMPUTER**

**DEPOK**

**JULI, 2009**

## HALAMAN PERNYATAAN ORISINALITAS

**Skripsi ini adalah hasil karya saya sendiri,  
semua sumber yang saya gunakan pada skripsi ini  
telah saya nyatakan dengan benar.**

Nama : Armando Yonathan

NPM : 1205000177

Tanda Tangan :

Tanggal : 21 Juli 2009

## HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Armando Yonathan

NPM : 1205000177

Program Studi : Ilmu Komputer

Judul Skripsi : Perolehan Informasi Dokumen Suara Pembicaraan Berdasarkan Hasil dari Sistem Penegenalan Suara untuk Bahasa Indonesia.

**Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana S.Kom pada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia**

### DEWAN PENGUJI

Pembimbing : Dra. Mirna Adriani Ph.D. (.....)

Penguji : Ika Alfina, S.Kom, M.Kom. (.....)

Penguji : Dadan Hardianto, S.Kom, M.Kom. (.....)

Ditetapkan di : Fakultas Ilmu Komputer

Tanggal : 21 Juli 2009

## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama : Armando Yonathan  
NPM : 1205000177  
Program Studi : Ilmu Komputer  
Fakultas : Ilmu Komputer  
Jenis karya : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif** (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul :  
Perolehan Informasi Dokumen Suara Pembicaraan Berdasarkan Hasil dari Sistem Penengenalan Suara untuk Bahasa Indonesia

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok  
Pada tanggal : 21 Juli 2009  
Yang menyatakan

( Armando Yonathan )

## KATA PENGANTAR

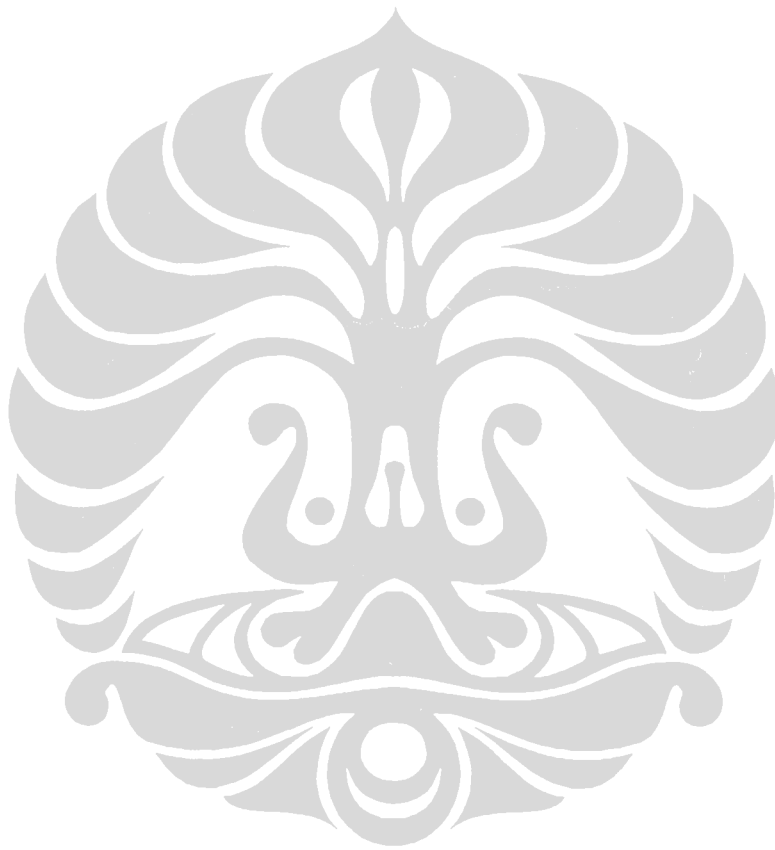
Penulis mengucapkan puji dan syukur kepada Tuhan Yang Maha Esa atas segala rahmat dan karunia yang telah diberikan sehingga penulis dapat menyelesaikan tugas akhir dan menghasilkan laporan tugas akhir ini dalam waktu yang telah ditentukan. Penulis juga mengucapkan terima kasih kepada semua pihak yang telah memberikan bantuan, dukungan, dan semangat kepada penulis selama pelaksanaan tugas akhir dan dalam penulisan laporan pelaksanaan tugas akhir ini. Penulis ucapkan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua penulis yang selalu mendoakan dan mendukung semua kegiatan yang dilakukan penulis.
2. Ibu Mirna Adriani selaku dosen pembimbing tugas akhir.
3. Ibu Dina Cahyati selaku pembimbing akademis.
4. Baskoro dan Ame yang telah mengajarkan cara penggunaan Sphinx dan memberikan penjelasan yang banyak tentang pengenalan suara otomatis bahasa Indonesia.
5. Bayu, Bernadia, Clara, Darwin, Deny, Hansel, Heninggar, Metti, Mulyandra, Refly, Rizal, Suryanto, Teddy, Vinky dan semua rekan Lab IR yang telah berjuang bersama penulis selama mengerjakan tugas akhir. Terima kasih untuk semua semangat dan nasihat yang sudah diberikan.
6. Adit, Bambang, Haryadi, Prajna, Rizky dan semua rekan-rekan futsal dan badminton.
7. Mursal yang telah dengan sukarela memberikan kritik tentang penulisan laporan tugas akhir ini.
8. Teman-teman Fakultas Ilmu Komputer Universitas Indonesia angkatan 2005.
9. Semua staf pengajar dan keluarga besar Fakultas Ilmu Komputer Universitas Indonesia.
10. Semua pihak yang tidak dapat ditulis satu-persatu yang telah membantu penulis selama pelaksanaan tugas akhir.

Penulis menyadari bahwa laporan ini masih memiliki kekurangan dan kesalahan. Karena itu, penulis mengharapkan saran dan masukan yang membangun untuk laporan ini. Semoga laporan ini dapat bermanfaat bagi siapa saja yang membacanya

Depok, 17 Juli 2009

Armando Yonathan



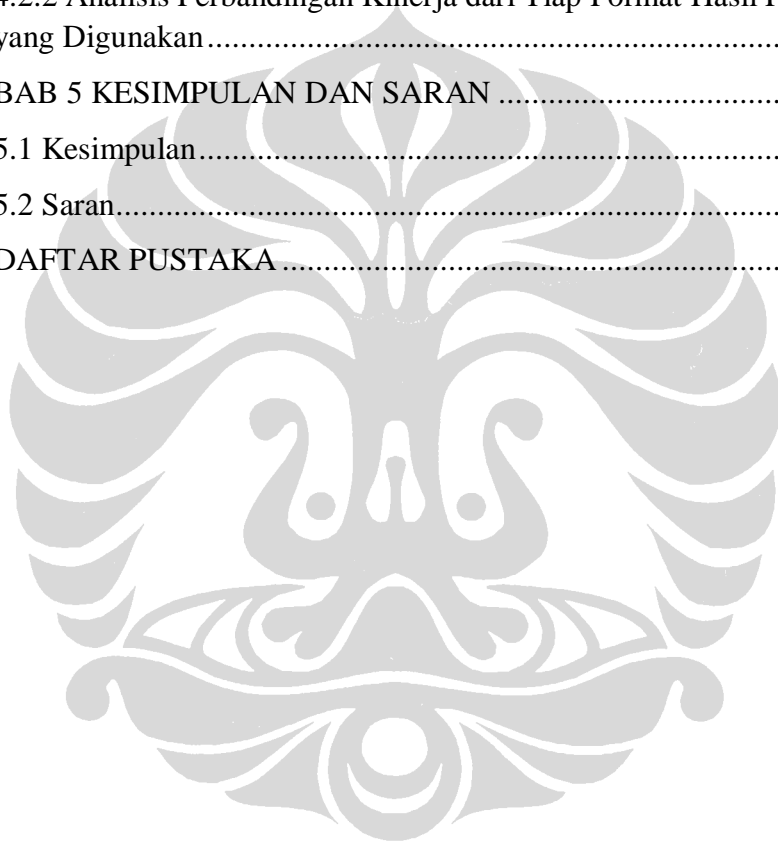
## DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERNYATAAN ORISINALITAS .....	ii
HALAMAN PENGESAHAN .....	iii
KATA PENGANTAR .....	v
ABSTRAK .....	vii
ABSTRACT .....	viii
DAFTAR ISI .....	ix
DAFTAR GAMBAR .....	xii
DAFTAR TABEL .....	xiii
DAFTAR LAMPIRAN.....	xiv
BAB 1PENDAHULUAN .....	1
1.1Latar Belakang.....	1
1.2Permasalahan .....	3
1.3Tujuan.....	3
1.4Ruang Lingkup .....	4
1.5Metodologi Penelitian .....	4
1.6Sistematika Penulisan.....	5
BAB 2LANDASAN TEORI.....	6
2.1 Perolehan Informasi .....	6
2.2 Sistem Perolehan Informasi.....	6
2.2.1 Tokenisasi.....	8
2.2.2 Pembuangan <i>Stopwords</i> .....	8
2.2.3 Pemotongan Imbuhan.....	9
2.2.4 Pembobotan Kata .....	9
2.2.5 Pengindeksan .....	11
2.2.6 Model Perolehan Informasi .....	12
2.2.7 Evaluasi .....	15
2.3 Perolehan Informasi Dokumen Suara Pembicaraan.....	16



2.4 Pengenalan Data Suara.....	18
2.5 Proses Pengenalan Suara .....	19
2.5.1 Model Akustik .....	21
2.5.2 Model Bahasa .....	22
2.5.3 Kamus Pelafalan dan Kamus Fonetik .....	23
2.6 Hasil Pengenalan Suara .....	25
2.6.1 <i>1-Best Output</i> .....	25
2.6.2 <i>N-Best Words Output</i> .....	25
2.6.3 <i>N-Best Pronunciation Output</i> .....	26
2.6.4 <i>Word Posterior Lattice</i> .....	26
2.7 Studi Literatur Penelitian Mengenai Perolehan Informasi Dokumen Suara Pembicaraan .....	28
2.7.1 Penggunaan <i>Position Specific Posterior Lattice</i> untuk Mengindeks Dokumen Suara Pembicaraan (Chelba, 2005).....	28
2.7.2.... Perolehan Informasi Dokumen Suara Pada Koleksi Berita Berbahasa India (Shah, 2001) .....	29
2.7.3.... Perolehan Informasi Dokumen Suara Pada Koleksi Perkuliahan Berbahasa Korea (Lee, 2008).....	30
<b>BAB 3 EKSPERIMEN</b> .....	29
3.1 Data Suara .....	29
3.2 Kamus Fonetik Bahasa Indonesia.....	30
3.3 Aplikasi yang Digunakan .....	31
3.3.1 Sphinx-4 .....	31
3.3.2 Indri 2.8 .....	33
3.3.3 Trec_Eval.....	33
3.3.4 CMU-LM.....	34
3.3.5 SphinxTrain1.0.....	35
3.4 Implementasi.....	36
3.4.1 Pelatihan Model Akustik .....	36
3.4.2 Penilaian Dokumen Relevan.....	37
3.4.3 Implementasi <i>1-Best Output</i> .....	38
3.4.4 Implementasi <i>N-Best Words Output</i> .....	39
3.4.5 Implementasi <i>N-Best Pronunciation Output</i> .....	41

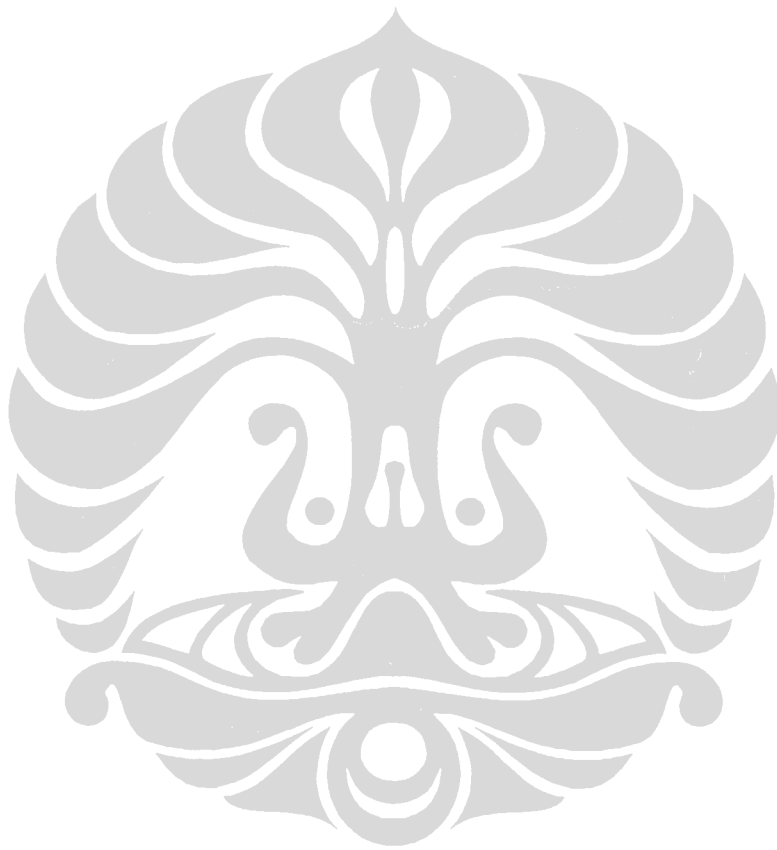
3.4.6 Implementasi <i>Word Posterior Lattice</i> .....	42
<b>BAB 4 HASIL EKSPERIMEN DAN ANALISIS</b> .....	<b>44</b>
4.1 Hasil Eksperimen .....	44
4.1.1 Hasil Eksperimen Berdasarkan Hasil Pengenalan Suara .....	44
4.1.2 Hasil Eksperimen Berdasarkan Jenis Kueri.....	46
4.2 Analisis .....	56
4.2.1 Analisis Perbandingan Kinerja dari Tiap Format Hasil Pengenalan Suara yang Digunakan pada Setiap Kueri .....	56
4.2.2 Analisis Perbandingan Kinerja dari Tiap Format Hasil Pengenalan Suara yang Digunakan.....	57
<b>BAB 5 KESIMPULAN DAN SARAN</b> .....	<b>61</b>
5.1 Kesimpulan.....	61
5.2 Saran.....	62
<b>DAFTAR PUSTAKA</b> .....	<b>63</b>



## DAFTAR GAMBAR

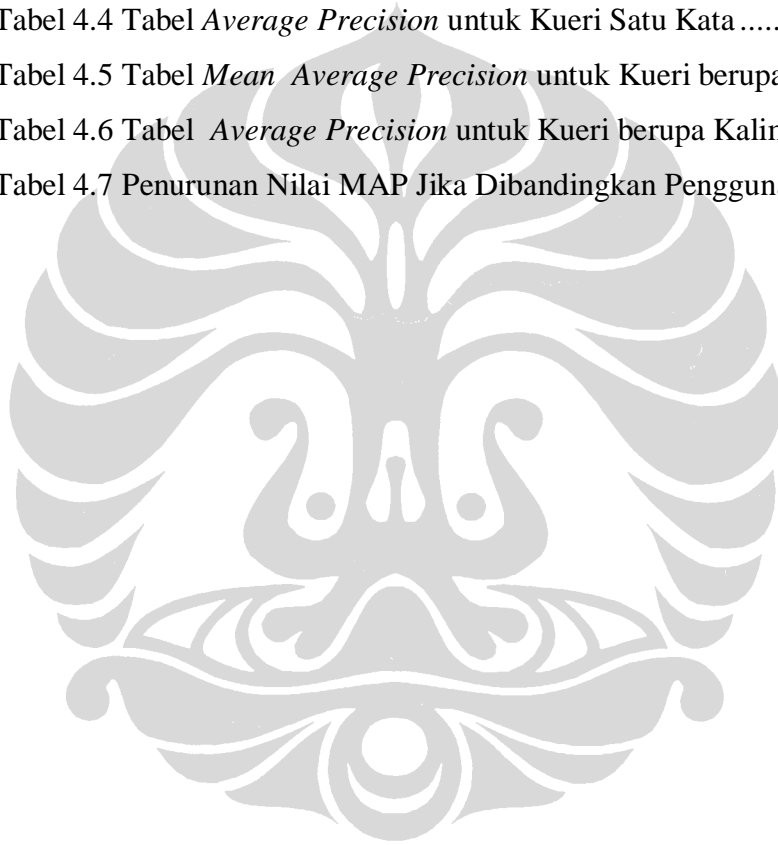
Gambar 2.1 Sistem Perolehan Informasi .....	7
Gambar 2.2 Proses Tokenisasi.....	8
Gambar 2.3 Jaringan Inferensia Dasar untuk Dokumen (Turtle, 1991).....	14
Gambar 2.4 Sistem Perolehan Informasi Dokumen Suara Pembicaraan.....	17
Gambar 2.5 Contoh Kamus Fonetik.....	24
Gambar 2.6 Contoh Kamus Pelafalan.....	24
Gambar 2.7 Representasi <i>Words Posterior Lattice</i> .....	27
Gambar 2.8 Proses Membuat <i>Contents Table dan Matching Table</i> (Lee, 2008).....	30
Gambar 3.1 Tahap Pembuatan Model Bahasa dengan CMU-LM (Clarkson, 1997).....	34
Gambar 3.2 Contoh Teks Transkripsi.....	36
Gambar 3.3 Contoh <i>Filler</i> .....	37
Gambar 3.4 Contoh Kamus Pelafalan untuk Proses Pelatihan.....	37
Gambar 3.5 Contoh Hasil Penilaian Dokumen Relevan.....	38
Gambar 3.6 Eksperimen <i>1-Best Output</i> .....	38
Gambar 3.7 Contoh Hasil Pengenalan Suara dengan Kemungkinan Terbaik.....	39
Gambar 3.8 Contoh Hasil Perolehan Dokumen.....	39
Gambar 3.9 Eksperimen <i>N-Best Words Output</i> .....	40
Gambar 3.10 Contoh Semua Kandidat Hasil Pengenalan Suara.....	40
Gambar 3.11 Contoh Bentuk Pelafalan Hasil Pengenalan Suara.....	41
Gambar 3.12 Eksperimen <i>N-Best Pronunciation Output</i> .....	41
Gambar 3.13 Contoh <i>Path</i> dari Lattice Hasil Pengenalan Suara.....	42
Gambar 3.14 Eksperimen <i>Word Posterior Lattice</i> .....	43
Gambar 4.1 Grafik MAP untuk Setiap Eksperimen.....	45

Gambar 4.2 Grafik Jumlah Dokumen Relevan yang Diperoleh Pada Setiap Eksperimen..	46
Gambar 4.3 Grafik <i>Average Precision</i> untuk Kueri Frase.....	49
Gambar 4.4 Grafik <i>Average Precision</i> untuk Kueri Satu Kata.....	52
Gambar 4.5 Grafik <i>Average Precision</i> untuk Kueri berupa Kalimat.....	55



## DAFTAR TABEL

Tabel 3.1 Konfigurasi Dokumen Suara Sphinx-4 .....	29
Tabel 3.2 Daftar Fonem pada Kamus Fonetik dan Contoh Pemakaiannya dalam Kata.....	30
Tabel 4.1 Tabel <i>Mean Average Precision</i> untuk Kueri Frase .....	47
Tabel 4.2 Tabel <i>Average Precision</i> untuk Kueri Frase.....	47
Tabel 4.3 Tabel <i>Mean Average Precision</i> untuk Kueri Satu Kata .....	50
Tabel 4.4 Tabel <i>Average Precision</i> untuk Kueri Satu Kata.....	50
Tabel 4.5 Tabel <i>Mean Average Precision</i> untuk Kueri berupa Kalimat.....	53
Tabel 4.6 Tabel <i>Average Precision</i> untuk Kueri berupa Kalimat .....	53
Tabel 4.7 Penurunan Nilai MAP Jika Dibandingkan Penggunaan <i>1-best output</i> .	58



## DAFTAR LAMPIRAN

LAMPIRAN 1 KUERI.....	66
LAMPIRAN 2 HASIL EVALUASI.....	80

