ISSN : 2085-4218

Text To Speech Bahasa Indonesia Menggunakan Metode Dhipone Concatenation

Ahmad Fahrudi Setiawan
Jurusan Teknik Infomatika Fakultas Teknik Industri ITN Malang, Jl. Raya Candi VIB Kav. Pairs No 2 Tidar Malang,
E-mail : fahrudi.itn@gmail.com

Abstrak. Telah dilakukan penelitian yang berjudul Aplikasi text to speech Bahasa Indonesia menggunakan metode metode *dhipone concatenation* yang bertujuan membangun software untuk konversi dari text ke suara yang dapat mengeluarkan suara dari inputan text yang disediakan. Program dapat digunakan untuk merubah tulisan menjadi suara.

Program ini dibangun dengan konsep desktop yang dapat digunakan dengan mudah karena hanya menyediakan 1 buah form. Di dalam program ini user dapat mencopy paste text kedalam program dengan cepat dan mendengarkan suara dari hasil text yang telah di salin.

Secara umum proses dalam sistem *text to speech* terdiri dari *dhipone concatenation* yang berupa modul konversi teks ke *fonem* yang menghasilkan transkripsi fonetik beserta informasi intonasi dan ritme (dikenal dengan prosodi) dan *Digital Signal Processing* (DSP) yang berupa modul konversi *fonem* ke ucapan yang mengubah informasi *fonetis* yang diterimanya menjadi sinyal ucapan.

Kata Kunci: *Text to speech, Dhipone concatenation*.

1. Pendahuluan

Membaca adalah jendela dunia, dengan membaca kita memasuki pintu ilmu pengetahuan. Para ilmuan membaca setiap hari untuk bisa mengerti, memahami dan menjadi pintar. Tetapi dengan kesibukan yang sangat padat maka membaca menjadi sulit dilakukan, kondisi ini memunculkan ide bagaimana dengan teknologi komputer manusia tetap beraktifitas seperti biasa tetapi dapat membaca dengan bantuan komputer untuk merubah text to speech. Dari kondisi ini manusia tetap bisa melakukan kegiatan yang lain dan tetap bisa belajar dengan program ini dan seolah-olah membaca dengan menggunakan *Audio-Learning*.

Audio-learning merupakan pembelajaran dengan memanfaat teknologi elektronik yang berupa suara, yaitu hanya memilih *digital book* apa yang ingin di dengarkan dengan hanya mengklik *digital book* tersebut maka kompjuter langsung membacakannya. Teknik yang digunakan adalah *diphone concatenation* dengan prinsip yang terdiri dari dua sub sistem, yaitu: Bagian kenverter text ke fonem (*text to phoneme*), serta Bagian konverter fonem to ucapan (*phoneme to speech*).

Bagian Konverter Teks ke Fonem berfungsi untuk mengubah kalimat masukan dalam suatu bahasa tertentu yang berbentuk teks menjadi rangkaian kode-kode bunyi yang biasanya direpresentasikan dengan kode fonem, durasi serta *pitch*-nya. Bagian ini bersifat sangat *language dependant*. Untuk suatu bahasa baru, bagian ini harus dikembangkan secara lengkap khusus untuk bahasa tersebut. Bagian Konverter Fonem ke Ucapan akan menerima masukan berupa kode-kode fonem serta pitch dan durasi yang dihasilkan oleh bagian sebelumnya. Berdasarkan kode-kode tersebut, bagian Konverter Fonem ke Ucapan akan menghasilkan bunyi atau sinyal ucapan yang sesuai dengan kalimat yang ingin diucapkan. Ada beberapa alternatif teknik yang dapat digunakan untuk implementasi bagian ini.

Dua teknik yang banyak digunakan adalah *formant synthesizer*, serta *diphone concatenation*. Formant *synthesizer* bekerja berdasarkan suatu model matematis yang akan melakukan komputasi untuk menghasilkan sinyal ucapan yang diinginkan. Synthesizer jenis ini telah lama digunakan pada berbagai aplikasi. Walaupun dapat menghasilkan ucapan dengan tingkat kemudahan interpretasi yang baik, *synthesizer* ini tidak dapat menghasilkan ucapan dengan tingkat kealamian yang tinggi.

Berdasarkan penelitian-penelitian yang telah dilakukan, maka dapat diasumsikan bahwa *Synthesizer* yang menggunakan teknik *diphone concatenation* dapat menghasilkan bunyi ucapan dengan tingkat kealamian (*naturalness*) yang tinggi.

ISSN : 2085-4218

2. Pengertian *Text To Speech*

Sistem konversi *text-to-speech* (TTS) merupakan suatu sistem yang mampu memproduksi sinyal ucapan secara otomatis melalui transkripsi grafem-ke-*fonem* untuk kalimat yang diucapkan. Perbedaan sistem TTS dengan *talking machine* biasa adalah keotomatisannya dalam mengucapkan kata-kata baru. Oleh karena itu TTS memungkinkan untuk diimplementasikan pada bidang aplikasi yang beragam seperti aplikasi sms bicara, buku digital dan pembaca email otomatis.

Dutoit dalam buku *"An Introduction to Text-to-Speech Synthesis"* [1] mendefinisikan *Text-to-Speech* sebagai *"production of speech by machines, by way of the automatic phonetization of the sentences to utter"*. Vainio dalam buku *"Artificial Neural Network Based Prodosy Models for Finsih Text-to-Speech Synthesis"* [8] menyatakan bahwa *"The task of Text to Speech system is to convert plain text into speech"*. Dalam bagian lainnya, Pelton menyatakan pula *"A very attractive advantage of text to speech is that any text can be read, vocabulary is not restricted to utterances that have been decided upon beforehand"*.

Speech FAQ, suatu situs Internet yang merangkum pendapat dari berbagai universitas, lembaga penelitian dan industri di bidang aplikasi ucapan, menyatakan bahwa *"Speech synthesis programs convert written input to spoken output by automatically generating synthetic speech. Speech synthesis is often referred to a Text-to-Speech conversion (TTS)" [9]*. Secara umum proses dalam sistem TTS terdiri dari *Natural Language Prossesing* (NLP) yang berupa modul konversi teks ke *fonem* yang menghasilkan transkripsi fonetik beserta informasi intonasi dan ritme (dikenal dengan prosodi) dan *Digital Signal Processing* (DSP) yang berupa modul konversi *fonem* ke ucapan, yang mengubah informasi fonetis yang diterimanya menjadi sinyal ucapan [4].

3. *Natural Language Prossesing* (NLP)

Modul NLP dapat diimplementasikan dengan beberapa solusi, yang sering diklasifikasikan sebagai *dictionary-based* dan *rule-based.* Solusi *dictionary-based* diimplementasikan dengan menyimpan sebanyak mungkin informasi *fonologi* ke dalam kamus [2]. Dalam metoda ini transkripsi dilakukan dengan cara metoda *lookup database* leksikal yang telah disusun. Sedangkan sistem transkripsi *rule-based,* menggantikan penyimpanan informasi fonologi dalam kamus dengan membuat set aturan *letter-to-sound* (atau *grafem*-ke-*fonem*).

3.1. *Synthesizer*

Tahap pemrosesan terakhir dari sistem TTS adalah sintesa sinyal ucapan. Secara umum terdapat tiga metoda dasar untuk sintesa sinyal ucapan. Sintesis *articulatory*, yang berusaha memodelkan sistem produksi sinyal ucapan manusia dengan pendekatan fisik mekanis secara langsung, *formant synthesizer* yang memodelkan frekuensi pole suatu sinyal ucapan atau fungsi transfer yang berbasis vocal track atau model source-filter, *synthesizer* perangkaian (*concatenation*), yang menggunakan panjang bagian yang berbeda dari suatu perekaman sinyal ucapan alami [5]. Namun demikian dua teknik yang sering digunakan adalah *formant synthesizer* dan *diphone concatenation*.

3.2. Fonem

Fonem dapat juga digunakan sebagai unit ucapan pada *database* [3]. Beberapa masalah pada sintesis perangkaian dibandingkan dengan metoda yang lain, yaitu:

1. Terjadi distorsi akibat ketidakberlanjutan pada titik perangkaian, yang dapat dikurangi dengan menggunakan *diphone* atau beberapa metoda lainnya untuk memperhalus sinyal ucapan.
2. Kebutuhan terhadap memori sangat tinggi, khususnya ketika menggunakan unit perangkaian yang panjang, misalnya suku kata dan kata.
3. Pengumpulan data dan penandaan bagian sinyal ucapan membutuhkan waktu yang lama.

ISSN : 2085-4218

3.3. Konversi dari Text to Speech

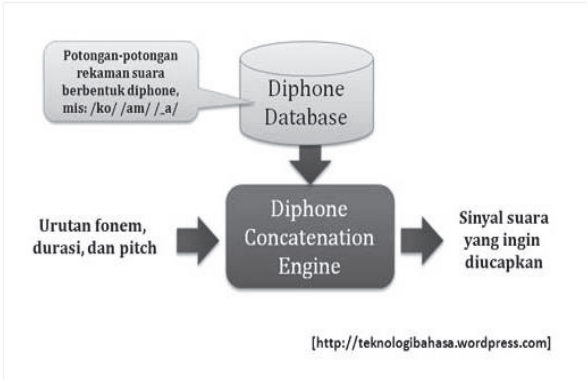
Text to Speech (TTS) diartikan sebagai proses pengubahan teks menjadi audio digital dan diucapkan. Pengucapan ini dapat berupa pengiriman audio digital tersebut ke pengeras suara computer atau menyimpan hasil pengubahan tersebut untuk diputar nanti. Tugas sistem TTS secara umum dapat dibagi dalam 2 bagian besar, analisa teks dan sintesa ucapan. Analisa teks mentransformasi teks masukan menjadi representasi linguistik, selanjutnya bagian sintesis ucapan mentransformasi representasi linguistik tersebut menjadi gelombang sinyal ucapan.

Sistem *Text to Speech* pada prinsipnya terdiri dari dua sub sistem, yaitu: Bagian Konverter Teks ke Fonem (*Text to Phoneme*), serta Bagian Konverter Fonem to Ucapan (*Phoneme to Speech*). Bagian Konverter *text* ke Fonem berfungsi untuk mengubah kalimat masukan dalam suatu bahasa tertentu yang berbentuk teks menjadi rangkaian kode-kode bunyi yang biasanya direpresentasikan dengan kode fonem, durasi serta *pitch*-nya [6].

Bagian Konverter Fonem ke Ucapan akan menerima masukan berupa kode-kode fonem serta pitch dan durasi yang dihasilkan oleh bagian sebelumnya. Berdasarkan kode-kode tersebut, bagian Konverter Fonem ke Ucapan akan menghasilkan bunyi atau sinyal ucapan yang sesuai dengan kalimat yang ingin diucapkan.

3.4. Diphone Concatenation

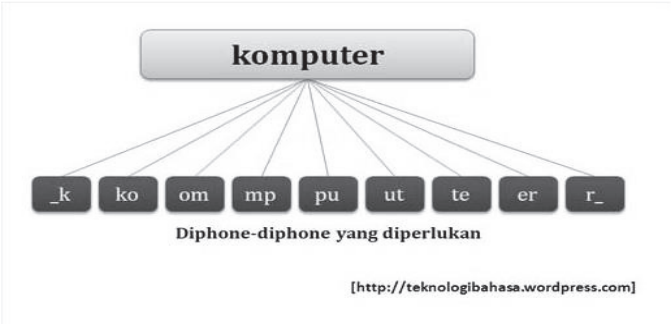
Teknik *diphone concatenation* bekerja dengan cara menggabung-gabungkan segmen-segmen bunyi yang telah direkam sebelumnya. Setiap segmen berupa *diphone* (gabungan dua buah fonem). *Synthesizer* jenis ini dapat menghasilkan bunyi ucapan dengan tingkat kealamian (*naturalness*) yang tinggi. Pembentukan ucapan pada pensintesa ucapan menggunakan metode *diphone concatenation* pada prinsipnya dilakukan dengan cara menyusun sejumlah *diphone* yang bersesuaian sehingga diperoleh ucapan yang diinginkan [7]. Sebagai contoh, pada gambar diperlihatkan pembentukan kata atau ucapan "komputer" yang disusun dari diphone-diphone /_k/, /ko/, /om/ dan seterusnya.



Gambar 1 Blok Diagram Pembentukan Ucapan "komputer" dari Diphone

Supaya pensitesa ucapan dapat mengucapkan semua kemungkinan kata atau kalimat yang ada dalam suatu bahasa, sehingga sistem tersebut harus didukung oleh *diphone database* yang terdiri dari semua kombinasi *diphone* yang ada dalam bahasa tersebut. *Diphone concatenation engine* atau unit pemroses *diphone* akan menerima masukan berupa daftar *fonem* yang ingin diucapkan, masing-masing disertai oleh durasi pengucapannya, serta *pitch* atau frekuensinya. Berdasarkan daftar fonem yang diterima, unit ini akan menentukan susunan *diphone* yang sesuai. Selanjutnya, unit ini akan melakukan smoothing sambungan antar *diphone*, melakukan manipulasi durasi pengucapan serta manipulasi *pitch* (lihat Gambar di bawah). Pada akhirnya, *diphone concatenation engine* akan menghasilkan sinyal ucapan yang sesuai.

ISSN : 2085-4218



Gambar 2 Pembentukan Ucapan "komputer" dari Diphone.

4. Implementasi Program

4.1 Form Utama

Tampilan utama dari aplikasi Audio Learning menggunakan metode *Diphone Concatenation* terlihat pada gambar 3. berikut ini:

Gambar 3 Tampilan Utama Aplikasi

Tampilan utama pada aplikasi merupakan *form* yang digunakan untuk tampilan awal dari program text to speech, terdapat berbagai menu yang bisa dipilih oleh pengguna. Dalam aplikasi ini memberikan kemudahan kepada pengguna dalam memilih menu yang telah disediakan, karena menu tersebut tersedia di *form* tampilan utama.

4.2 Form Pilih File

Pada pilihan *menu* Pilih file merupakan tombol yang akan mengarahkan ke *menu open*. Menu ini akan menampilkan pilihan file tersimpan berekstensi *.PDF dan *.TXT yang ingin anda buka.

ISSN : 2085-4218

Gambar 4 Tampilan Pilih File

Pada menu PDF merupakan pilihan yang mengarahkan pada tampilan yang berekstensikan .PDF, pengguna bisa membuka langsung file yang berektensi *.PDF pada pilihan menu PDF. Kemudian bisa mengcopykan paragraph yang akan diproses menjadi *speech* ke menu SPEAKING. Pada menu Speaking merupakan menu yang menampilakan teks-teks yang diproses menjadi *speech*. Ketikan teks-teks tersebut telah diproses menjadi *speech* pengguna bisa langsung mendegarkan isi yang terdapat dalam teks tersebut.

5. Form *Text To Speech*

Text to speech dirancang menggunakan bahasa Indonesia memudahkan pengguna yang mengerti berbahasa Indonesia. Karena *prosody* di setiap bahasa berbeda-beda, ketika program *text to speech* menggunakan bahasa inggris atau *prosody* inggris tetapi isi dari teks mengguakan bahasa Indonesia maka speech yang akan di dengar adalah bahasa Indonesia yang berprosody inggris. Hasil tersebut akan membingungkan pengguna untuk memahami isi teks tersebut.

Ketika proses speech berlangsung terdapat status *speaking* yang menunjukkan bahwa teks sedang dibacakan. Apabila proses speech telah selesai maka status yang ditunnjukkan yaitu diam, dan status *not speaking* menunjukkan bahwa dalam memo tidak terdapat teks yang akan di proses menjadi speech. Speech yang dibacakan bisa diatur kecepatannya dalam mebacakan teks dengan mengatur *trackbar* yang sudah tersedia dalam aplikasi ini. Pada tombol katakana dan stop merupakan tombol untuk proses *speech* dan menghentikan *speech*.

ISSN : 2085-4218

Gambar 5 Tampilan *Text to Speech*

6. Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah:

1. Program dapat membaca tulisan yang diinputkan penguna dengan benar.
2. Program dapat membaca tanda baca.
3. Program dapat melafalkan ucapan bahasa Indonesia dengan cukup baik, walaupun tidak sebaik manusia.

7. Daftar Pustaka

[1] Dutoit, Thierry (1996), High-Quality Text-to-Speech Synthesis : an Overview, Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis, vol. 17 n°1, pp. 25-37.
[2] Haykin, Simon (1994). Neural Network : A Comprehensive Foundation, Macmillan Publishing Company.
[3] Moulines E., F. Charpentier (1990), Pitch Synchronous Waveform Processing technique for text-to-speech Synthesis Using Diphones, Speech Comunication, August, 453-467.
[4] JR, John R. Deller, John G. Proakis, John H.L. Hansen (1993), Discrete-Time Processing of Speech Signals, Macmillan Inc.
[5] Rabiner, R. Schafer (1978), Digital Signal Processing of Speech Signal, Signal Processing, Prentice Hall.
[6] Shaw-Hwa Hwang and Sin-Horng Chen (1995), "A Prosodic Model of Mandarin Speech And Its application to Pitch Level Generation for Text-to-Speech", IEEE.
[7] Sin-Horng Chen, Shaw-Hwa Hwang and Yih-Ru Wang (May 1998),"An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE, Vol 6 No.3
[8] Vainio, Marti (2001), "Artificial Neural Network Based Prodosy Models for Finsih Text-to-Speech Synthesis", University of Helsinki.
[9] Sagisaka,"On the prediction of Global F0 shape for Japanese text-to-speech" ICASSP, pp.325-328, 1990
[10] Www.speech.cs.cmu.edu/comp.speech/index.html dengan penanggung jawab Andrew Hunt dari Speech Applications Group, Sun Micro systems Laboratories.