PEMBANGKITAN *PROSODY* PADA *TEXT TO SPEECH SYNTHESIS SYSTEM* UNTUK PENUTUR BERBAHASA INDONESIA

Zonda Rugmiaga[1], Miftahul Huda[2]
Mahasiswa Jurusan Teknik Telekomunikasi[1], Dosen Pembimbing[2]
Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember (ITS) Surabaya
Kampus PENS-ITS, Keputih, Sukolilo, Surabaya.
Telp : (+62)31-5947280; Fax. (+62)31-5946011
Email : Zonda@student.eepis-its.edu

Abstrak

Text-To-Speech Synthesis System adalah sebuah konverter tulisan menjadi sebuah ucapan/audio*(spoken language)* yang bisa di dengar oleh user. Proses untuk membuat *Text-To-Speech Synthesis System* ini terdapat tiga tahap,yaitu *pre-text processing,prosody* dan *concatenation*. Khusus di dalam proses *prosody* ini,terdapat beberapa tahap yang harus di lakukan. Yaitu *The Multi Level Data Structure, Diphone Retrieval* dan *Accouctic Manipulation.* Melalui proses *prosody* di setiap ujung-ujung persambungan akan dilakukan proses PSOLA untuk memperhalus transisi antar sinyal *diphone*.

Pada pengujiannya menggunakan *pitch countour* didapatkan hasil bahwa *overlap* 30% memiliki *Pitch countour* yang lebih bagus dengan jumlah lebih sedikit sinyal *drop* pada sambungannya bila di bandingan dengan penyambungan menggunakan *overlap* 50% dan 70%. Pengujian dengan survey kepada 20 responden , *overlap* 30%, mendapatkan nilai MOS 2.85 untuk *overlap* 30%, *overlap* 50% mendapatkan nilai MOS 2.81 dan *overlap* 70% mendapatkan nilai MOS 2.79. sehingga,penggunaan *overlap* 30% lebih bagus hasilnya bila dibandingkan dengan menggunakan *overlap* 50% dan 70%.

1. Pendahuluan

Text to speech adalah sebuah konverter yang bisa mengubah tulisan menjadi suatu *audio* yang bisa di dengar. Bahasa merupakan alat komunikasi paling tepat dalam melakukan pendekatan yang efektif untuk menyampaikan dan memahami ekspresi,keinginan dan maksud manusia. Bentuk representasinya adalah baik itu tulisan berupa suara atau ucapan *(spoken language)*.

Konverter ini banyak digunakan untuk membantu mempermudah manusia yang memiliki kekurangan seperti tuna wicara atau tunanetra, bahkan konverter ini juga bisa digunakan untuk mempermudah memberi pengajaran kepada anak-anak atau orang-orang yang masih belajar tulis menulis dan membaca. Namun dalam perkembangannya, konverter ini juga banyak diintegrasikan pada aplikasi-aplikasi lain. Seperti pada *handphone* yang memiliki OS symbian. Konverter ini juga digunakan sebagai *message reader*.

Dalam proyek akhir ini, akan dilakukan perancangan sistem yang mengkonversikan sebuah teks bahasa Indonesia ke dalam bentuk ucapan yang kemudian di jadikan *database* suara. *Text to speech synthesis system* meliputi : proses *text pre-processing*, pembangkitan *prosody* dan proses *concatenation* yang menggabungkan *diphone* dari *database* suara.

2. Dasar Teori

teori yang menjadi pembuatan system ini adalah:

2.1 Teknologi Pemrosesan Bahasa

Suatu sistem pemrosesan bahasa alami secara lisan dapat dibentuk dari tiga sub-sistem, yaitu sebagai berikut :

1. Sub-Sistem *Natural Language Processing* (NLP), berfungsi untuk melakukan pemrosesan secara simbolik terhadap bahasa tulisan. Beberapa bentuk aplikasi sub-sistem ini adalah translator bahasa alami (misalnya dari bahasa Inggris ke Bahasa Indonesia), sistem pemeriksaan sintak bahasa, sistem yang dapat menyimpulkan suatu narasi, dan sebagainya.
2. Sub-Sistem *Text-to-Speech* (TTS), berfungsi untuk mengubah teks (bahasa tulisan) menjadi ucapan (bahasa lisan).
3. Sub-Sistem *Speech Recognation* (SR), merupakan kebalikan teknologi *Text to speech*, yaitu sistem yang berfungsi untuk mengubah atau mengenali suatu ucapan (bahasa lisan) menjadi teks (bahasa tulisan).

2.2 Kaidah Bahasa Indonesia

Bahasa Indonesia mengenal bahasa tulisan maupun bahasa lisan. Dalam bahasa lisan, dikenal istilah fonem, yang merupakan kesatuan bahasa terkecil yang dapat membedakan arti. Dalam

bahasa tulisan, fonem dilambangkan dengan huruf. Seringkali istilah fonem disamakan dengan huruf, padahal tidak selamanya berlaku demikian.

2.1.1 Abjad

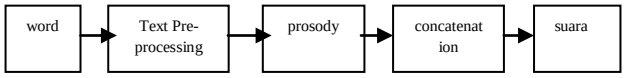
Abjad atau huruf yang digunakan dalam bahasa Indonesia terdiri atas 52 huruf, yaitu 26 huruf besar (A-Z) dan 26 huruf kecil (a-z).

2.1.2 Fonem

Fonem adalah istilah linguistik dan merupakan satuan terkecil dalam sebuah bahasa yang masih bisa menunjukkan perbedaan makna.

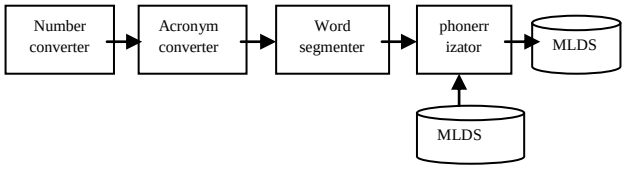
2.3 *Text-to-Speech Synthesis System*

Text to Speech synthesis system terdiri dari 3 bagian, yaitu *text pre-processing*, pembangkitan *prosody* dan *concatenation*. Diagram blok *text to speech synthesis system* adalah :



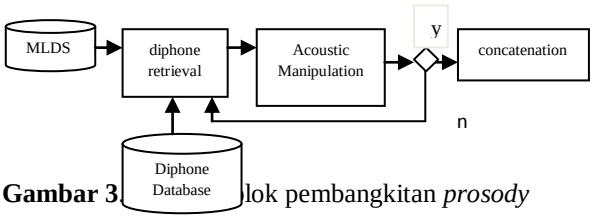
gambar 1. Diagram blok *text-to-speech synthesis system*

Pada bagian *Text Pre-processing*, terjadi pengkonversian dari *input* yang berupa *text* menjadi *diphone* (gabungan dua buah fonem). Ketika input yang berupa teks, akronim (singkatan) ataupun angka maka bagian ini akan mengkonversikan menjadi diphone yang telah tersedia di database *diphone*. Diagram blok untuk proses *text pre-processing* adalah :

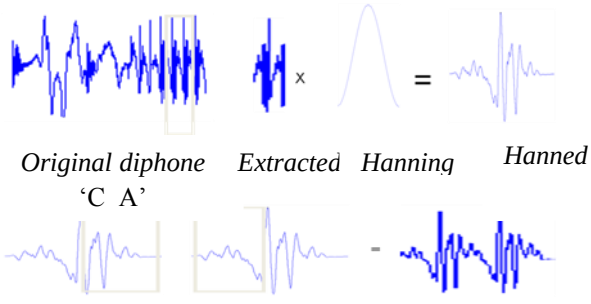


Gambar 2. Diagram blok *text pre-processing*

Pada proses pembangkitan *prosody* sangat memperhatikan karakter sinyal ucapan manusia,untuk mendapatkan ucapan yang lebih alami. Secara kuantisasi, prosodi adalah perubahan nilai *pitch* (frekuensi dasar) selama pengucapan kalimat dilakukan atau *pitch* sebagai fungsi waktu. Pembangkitan *prosody* ini bertujuan untuk memperhalus hasil proses *concatenation*. Jadi proses penggabungan *diphone-diphone* ini bisa menghasilkan suara yang mendekati naturalnya. Diagram blok untuk pembangkitan *prosody* adalah:



Gambar 3. Blok pembangkitan *prosody*



50% Overlap + Add

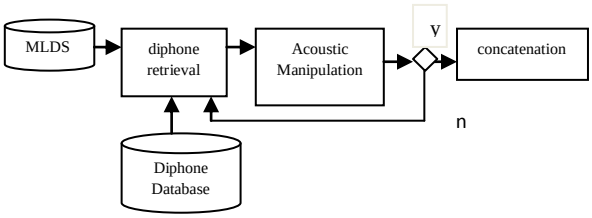
Gambar 4. Perubahan *Pitch* dengan PSOLA(sumber: Text-To-Speech Synthesis, Michael Beddaoui)

Concatenation Yaitu penggabung-gabungan segmen-segmen bunyi yang telah direkam sebelumnya. Setiap segmen berupa *diphone* (gabungan dua buah fonem). Pada perekaman suara dilakukan beberapa kali agar mendapatkan hasil yang akurat.

3. PERANCANGAN DAN PEMBUATAN SISTEM

3.1 Perencanaan blok diagram sistem

Berikut ini merupakan blok diagram pembangkitan Prosodi:



Gambar 5. Diagram blok system

Penjelasan dari blok tersebut adalah:

- MLDS (*The Multi_Level Data Structure*) tediri dari semua data yang diperlukan untuk sub_sistem berikutnya.
- *Diphone Retrieval* didalamnya ada tiga tahapan yang terjadi yait*u database* perekaman *diphone*,setia*p diphone* di matchkan dengan txt files (di bedakan oleh tipe CC,CV,VC,VV dan di referensikan ke komponen yang spesifik dalam bentuk gelombang), menyimpan bentuk gelomban*g diphone*
- *Diphone database* adalah direktori untuk menyimpan *diphone*
- *Accoustic Manipulation* di dalamnya terdapat proses pengenalan file-file gelombang .WAV(load,play,write) dan pemerosesan *pitch*..
- *Concatenation* adalah proses penyambungan *diphone*

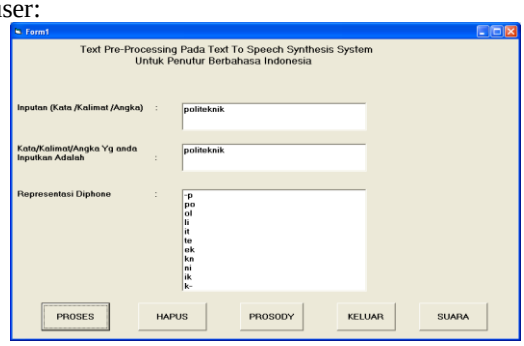
3.2 Pembuatan Sistem

Langkah-langkah pembuatan sistem secara umum:

1. Pembuatan program dengan Visual Basic 6 untuk mengkoneksikan Visual Basic 6 dengan Matlab
2. Pembuatan program dengan menggunakan Matlab untuk:
   a. Pendeteksian nilai *pitch* pada ujung frame yang paling awal dan yang paling akhir dari suatu *diphone* dan nambahkannya lagi ke *diphone* tersebut.
   b. Proses *overlap* antar satu *pitch* yang berada pada sisi transisi antar *diphone* dalam proses *concatenation*.
3. Pada metode PSOLA ada beberapa bagian yang menjadi prinsip dasar:
   a. Jika sinyal wav tersebut dalam proses *concatenation* terletak di paling depan,maka hanya satu *pitch period* yang paling terakhir itulah yang akan di proses.
   b. Jika sinyal vaw tersebut dalam proses *concatenation* terletak di tengah,maka ujung yang paling depan dan yang paling akhirlah yang akan diproses.
   c. Jika sinyal wav tersebut dalam proses *concatenation* paling belakang, maka hanya satu *pitch period* yang paling awal itulah yang akan di proses.

3.2.1 Pembuatan program dengan menggunakan vb6

Membuat *form* sebagai *interface* dengan user:



Gambar 6. Bentuk GUI *Text To Speech Synthesis System*

Selanjutnya pada tombol prosodi,di tambahkan program VB untuk memanggil file sukses.m yang di dalamnya terdapat program untuk melakukan proses PSOLA

3.2.2 Pembuatan program dengan menggunakan Matlab

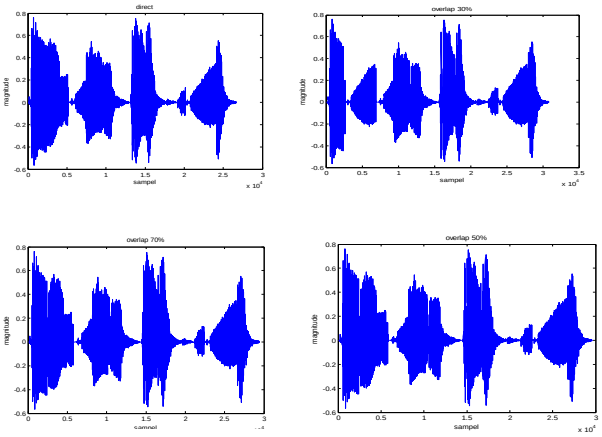
Pendeteksian nilai *pitch* menggunakan metode autokorelasi dengan frkuensi sampling 16000hz. Setelah diketahui nilai *pitch*,selanjutnya bisa di dapatkan panjang data dari *pitch* dengan rumus:

$$panjang\ data = \frac{1}{pitch} \times frek.sampling$$

Panjang data tersebut untuk menjadi patokan dalam penentuan banyak data yang akan di tambahkan ke *diphone* aslinya dan di *overlap*-kan.

Proses *overlap* di awali dengan membandingkan panjang data yang akan di *overlap*-kan dari masing-masing *pitch* dari kedua *diphone* yang akan digabungkan. Setelah itu akan di bandingkan dan dicari mana yang paling pendek. Panjang data yang paling pendek dari keduanya akan di jadikan patokan dalam penentuan banyak data yang akan di *overlap*-kan

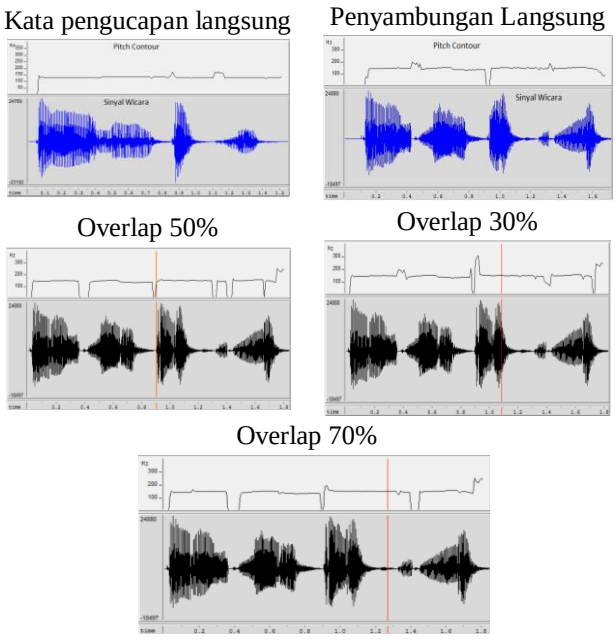
4. Hasil dan Analisa



Gambar 7. Hasil penyambungan *diphone-diphone* yang membentuk kata" politeknik" secara *direct* /langsung,psola overlap 30%,50% dan 70%

penyambungan dengan PSOLA membuat panjang data hasil penyambungan *diphone* tersebut lebih panjang bila di bandingkan dengan panjang data hasil penyambungan dengan metode langsung. penyambungan secara langsung memiliki panjang data keseluruhan 26661 sample,sedangkan pada penyambungan dengan menggunakan PSOLA,terdapat tiga pengkondisian. Yaitu menggunakan *overlap* 30%,50% dan 70%. Pada penyambungan dengan menggunakan metode PSOLA overlap 30% panjang data penyambungan sinyal *diphone-diphone* yang membentuk kata "politeknik" menjadi sepanjang 30832 sample,jika menggunakan PSOLA *overlap* 50%,panjang datanya menjadi 29598 sample. Dan jika menggunakan PSOLA dengan overlap 70%,panjang datanya sepanjang 29454 sample.

Dari ketiga metode PSOLA dengan besar *overlap* yang berbeda terlihat bahwa semakin panjangnya data yang akan di *overlap*-kan maka menyebabkan panjang data dari keseluruhan hasil penyambungan akan semakin pendek,hal ini di karenakan dua data yang di *overlap*-kan itu menjadi satu kesatuan atau di lebur menjadi satu,sehingga dalam hasil penyambungan memiliki panjang *diphone* sebesar panjang kedua *diphone* ditambah sisa kedua panjang data *pitch* yang tidak ikut di *overlap* kan dan panjang data yang telah di *overlap*-kan. Hal ini juga membuat *magnitude* dari data yang di overlapkan juga semakin besar.



Gambar 8 *Pitch countour* penyambungan kata-kata "politeknik" dengan pengucapan langsung,penyambungan langsung,PSOLA overlap 30%,50% dan 70%.

Pitch countour adalah fungsi atau kurva yang melacak *pitch* suara dari waktu ke waktu. Dari pengamatan yang kami lakukan dengan menggunakan *wavesurfer*. Dari hasil pengamatan pada sinyal wicara"Politeknik" secara langsung memiliki *Pitch countour* yang halus tanpa ada sinyal drop. Selanjutnya,hasil pengamatan pada sinyal wicara"Politeknik" yang di sambung secara langsung,memiliki *Pitch countour* yang lumayan halus dengan satu yang sinyal *drop*.

Setelah itu dengan metode PSOLA dengan overlap 30%,pada *pitch countour* terdapat tiga sinyal *drop* sempit. Pada PSOLA dengan overlap 50%,pada *pitch countour* terdapat tiga sinyal *drop*. Pada PSOLA dengan overlap 70%,pada *pitch countour* terdapat tiga sinyal *drop* tapi memiliki luasan yang lebih besar di bandingkan yang 50%. Sinyal *drop* itu di hasilkan karena perubahan mendadak dalam *pitch* atau *pitch* yang naik atau turun dari waktu ke waktu.

5. Kesimpulan

Berdasarkan hasil pembangkitan *Prosody* pada sisi penyambungan antar *diphone* yang telah dilakukan,ada beberapa kesimpulan yang bisa diambil:

1. Semakin besar tingkat *overlap pitch*-nya,maka semakin banyak sinyal *drop*-nya dan kualitas penyambungannya semakin jelek.
2. Penyambungan dengan metode PSOLA belum bisa memperbaiki proses penghalusan di sisi sambungan antar *diphone*.

6. Daftar Pustaka

[1] Beddaoui, Michael dan Aziz El-Solh, Abdel, *A Text To Speech Synthesis System*, 2002.

[2] Arman, Ary Akhmad, *Konversi Dari Teks Ke Ucapan*.pdf

[3] Silvia, Dina., *Text To Speech Pada PC Dengan Metode Direct Generation (Pembuatan Kaidah Pembangkitan Sinyal)*, PENS-ITS, Surabaya, 2005.

[4] Pusat Bahasa Departeman Pendidikan Nasional, *Kamus Besar Bahasa Indonesia*, Jakarta, 2008.

[5] Santoso, Tri Budi., Huda, Miftahul., Dutono, Titon., *Petunjuk Praktikum Aplikasi Pengolahan Sinyal Digital*, PENS, Surabaya, 2008.

[6] Yuswanto, *Microsoft Visual Basic 6.0*, PT Prestasi Pustaka, Surabaya, 2003.

[7] Komputer Wahana, *Pemrograman Visual Basic 6.0*, Penerbit Andi Yogyakarta, Semarang, 2000.