TEXT PRE-PROCESSING PADA TEXT TO SPEECH SYNTHESIS SYSTEM UNTUK PENUTUR BERBAHASA INDONESIA

Handi Dwi Rachma Bayu, Miftahul Huda
Jurusan Telekomunikasi, Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh November (ITS) Surabaya
Email : handee_eepis@yahoo.com

ABSTRAK

Perkembangan teknologi telekomunikasi yang sangat pesat menghasilkan berbagai produk teknologi telekomunikasi yang sangat beragam. Produk telekomunikasi yang dihasilkan dari waktu ke waktu merupakan produk yang lebih efektif dan efisien dalam hal penggunaan dan pemeliharaan, baik secara teknis maupun biaya.

Proyek akhir ini bertujuan untuk mengkonversikan penulisan teks pada PC atau laptop menjadi *output* suara sesuai dengan teks yang dituliskan. Pembuatan ini dilakukan dengan menggunakan metode *synthesis system* yang terdiri dari tiga proses yaitu *text pre processing*, pembangkitan *prosody* dan proses *concatenation.* Setelah diimplementasikan, perangkat lunak ini diuji coba sesuai dengan spesifikasi kebutuhan dan kemampuan yang dimiliki yaitu melakukan pengkonversian dari masukan kata atau kalimat ke bentuk representasi *diphone* yang kemudian *diphone-diphone* tersebut akan disambungkan (*concatenation*) untuk menjadi suara seperti teks yang diinputkan. Dengan demikian aplikasi perangkat lunak ini dapat digunakan untuk membantu para tuna netra agar bisa membaca berita dari internet ataupun membaca *email.*

Hasil dari paper ini adalah bahwa pada proses *text pre-processing* telah dapat mengkonversikan masukan yang berupa angka, akronim, kata atau kalimat ke bentuk representasi diphone.

Kata Kunci : *diphone, text pre-processing, prosody*

1. PENDAHULUAN

Perkembangan teknologi komputer yang sangat pesat, memicu perkembangan di berbagai bidang. Komputer diharapkan mampu berinteraksi secara lisan dengan pemakainya menggunakan bahasa sehari-hari, bukan bahasa mesin yang terkesan rumit.

Tentunya, komputer harus dilengkapi dengan perangkat lunak untuk mengendalikan semua sistem serta menjalankan fungsi-fungsinya. Perangkat keras serta sebagaian perangkat lunak akan bersifat generik, tetapi sebagian komponen perangkat lunaknya akan bersifat *language dependent*, yaitu perangkat lunak yang melakukan pemrosesan bahasa alami secara lisan.

Teknologi bahasa adalah teknologi yang berhubungan dengan penggunaan bahasa, baik bahasa lisan maupun bahasa tulisan. Bahasa merupakan alat komunikasi paling relevan dan tepat sasaran untuk menyampaikan keinginan dan maksud manusia. Bentuk representasinya adalah berupa suara atau ucapan *(spoken language)*, tetapi sering pula dinyatakan dalam bentuk tulisan. Sistem pemrosesan bahasa alami secara lisan dapat dibentuk dari sistem *text to speech.*

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

2. TEORI PENUNJANG

Teori yang digunakan sebagai dasar pembuatan sistem ini adalah adalah:

2.1 TEKNOLOGI PEMROSESAN BAHASA

Bahasa dapat dibedakan menjadi 2 , yaitu Bahasa Alami dan Bahasa Buatan. Bahasa alami adalah bahasa yang biasa digunakan untuk berkomunikasi antar manusia, misalnya bahasa Indonesia, Sunda, Jawa, Inggris, Jepang, dan sebagainya. Bahasa Buatan adalah bahasa yang dibuat secara khusus untuk memenuhi kebutuhan tertentu, misalnya bahasa pemodelan atau bahasa pemrograman komputer.

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

a. 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 sintaks bahasa, sistem yang dapat menyimpulkan suatu narasi, dan sebagainya.

b. Sub-Sistem *Text-to-Speech* (TTS), berfungsi untuk mengubah text (bahasa tulisan) menjadi ucapan (bahasa lisan).

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. Kadangkala terdapat beberapa perbedaan dalam kedua jenis bahasa ini. Dalam bahasa lisan, dikenal istilah fonem, yang merupakan kesatuan bahasa terkecil yang dapat membedakan arti. Dalam bahasa tulisan, fonem dilambangkan dengan huruf. Dengan kata lain, huruf adalah tulisan dari fonem. Seringkali istilah fonem disamakan dengan huruf, padahal tidak selamanya berlaku demikian. Berikut adalah konsep bahasa Indonesia berdasarkan pedoman umum ejaan bahasa Indonesia yang disempurnakan.

2.2.1 Abjad

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

2.2.2 Fonem

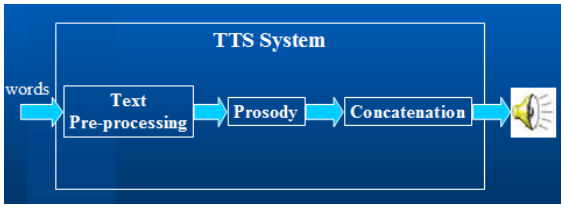
Fonem adalah istilah linguistik dan merupakan satuan terkecil dalam sebuah bahasa yang masih bisa menunjukkan perbedaan makna. Untuk bahasa Indonesia memiliki 35 fonem.

2.2.3 Diphone

Diphone adalah gabungan dari dua buah fonem bahasa Indonesia. Jumlah diphone dalam bahasa Indonesia kurang lebih sebanyak 1024 diphone.

2.3 TEXT TO SPEECH SYNTHESIS SYSTEM

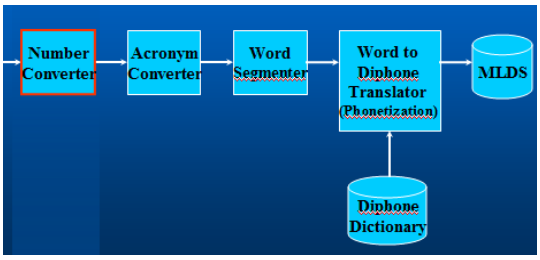
Text to Speech synthesis system terdiri dari 3 bagian, yaitu *text pre-processing*, pembangkitan *prosody* dan *concatenation*. Di bawah ini adalah diagram blok *text to speech synthesis system* :



Gambar 1. Blok Diagram Text to speech synthesis system [1]

2.3.1 Text pre-processing

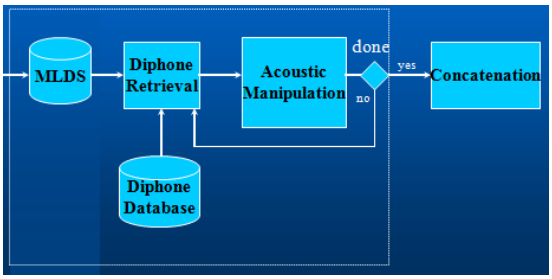
Yaitu pengkonversian dari input yang berupa teks menjadi *diphone* (gabungan dua buah fonem). Ketika masukan yang berupa teks, akronim (singkatan) ataupun angka maka bagian ini akan mengkonversikan menjadi diphone yang telah tersedia di database *diphone*. Diagram blok untuk proses *text pre-processing*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



Gambar 2. Blok diagram text pre processing [1]

2.3.2 Prosody

Yaitu untuk mendapatkan ucapan yang lebih alami, ucapan yang dihasilkan harus memiliki intonasi (*prosody*). Secara kuantisasi, prosodi adalah perubahan nilai *pitch* (frekuensi dasar) selama pengucapan kalimat dilakukan atau pitch sebagai fungsi waktu. Prosodi bersifat sangat spesifik untuk setiap bahasa, sehingga model yang diperlukan untuk membangkitkan data-data prosodi menjadi sangat spesifik juga untuk suatu bahasa. Diagram blok untuk prosodi adalah :



Gambar 3. Blok diagram Prosody [1]

- MLDS (*Multi Level Data Structure*), tediri dari semua data yang diperlukan untuk sub_sistem berikutnya. MLDS terdiri atas kata, reperesentasi *diphone, Prosodic* parameter untuk tiap diphone (ini perpaduan antara level kata dan level *prosody* kalimat). MLDS mengizinkan untuk modulasi.
- Diphone Retrieval didalamnya ada tiga tahapan yang terjadi, yaitu database perekaman *diphone*, setiap *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 gelombang *diphone* dan *Prosodic* parameter dalan variabel.

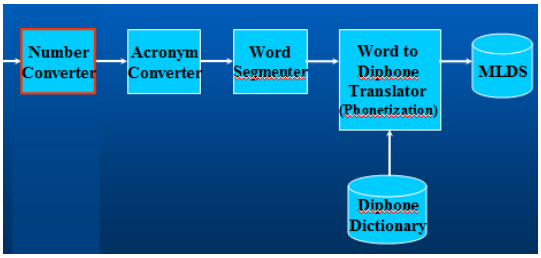
- Accoustic Manipulation di dalamnya terdapat proses pengenalan file-file gelombang .WAV(load, play, write), *vast array* dari peralatan *signal processing, built-in function, ease debugging , GUI-capable*

2.3.3 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 SISTEM

3.1 Perencanaan blok diagram sistem



Gambar 4. Blok diagram sistem [1]

Dari blok diagram sistem dapat dijelaskan cara kerja sistem yaitu :

- **Number Converter**
  Jika masukan pada sistem berupa angka, maka sistem mengkonversikan angka ke dalam representasi *diphone* (gabungan dua buah fonem).
- **Acronym Converter**
  Jika masukan pada sistem berupa kata singkatan dalam bahasa Indonesia, maka sistem mengkonversika singkatan ke dalam representasi *diphone* (gabungan dua buah fonem).
- **Word Segmenter**
  Jika masukan pada sistem berupa kata atau kalimat maka sistem mengkonversikan kata atau kalimat ke dalam representasi *diphone* (gabungan dua buah fonem).
- **Diphone Dictionary**
  Merupakan database yang berupa kumpulan dari *diphone – diphone.* Pembuatan diphone dilakukan dengan melakukan pelabelan pada kata. Jumlah *diphone* yang telah terkumpul sebanyak 394 *diphone*.
- **MLDS (Multi Level Data Structure)**
  Terdiri dari semua data yang diperlukan untuk sub system berikutnya, dalam hal ini adalah proses *prosody.* MLDS terdiri dari representasi diphone-diphone hasil pengkonversian inputan.

3.2 Implementasi Sistem

Langkah – langkah dalam pembuatan sistem secara umum adalah sebagai berikut :

1. Pembuatan *diphone database*
2. Pembuatan program menggunakan Microsoft Visual Basic 6.0 yang meliputi :
   - Pembuatan program pengkonversian dari masukan berupa angka, akronim, kata, dan kalimat ke representasi *diphone.*
3. Pembuatan program *direct concatenation* (Penyambungan langsung antar diphone) menggunakan Matlab 7.1

4. PENGUJIAN DAN ANALISA

4.1 Pembuatan Diphone Database

Pengujian pada diphone database dilakukan dengan cara melakukan survey kepada beberapa responden terhadap kualitas suara dari beberapa diphone secara acak yaitu po, ag, ad, at, li, me, sa, dw, gr, hw, hk, u-, -u, s-, dan –s. Dan dari pengujian tersebut didapatkan nilai *Mean Opinion Score* (MOS). Responden pada pengujian ini sebanyak 25 orang yang dipilih secara acak. Berikut merupakan tabel skala penilaian survey kuisioner MOS untuk file yang telah direkam :

Tabel 1. Skala penilaian MOS untuk file suara

Skala	Kualitas	Keterangan
5	Excellent	Sangat Jelas dan sangat jernih
4	Good	Jelas dan jernih
3	Fair	Cukup jelas dan cukup jernih
2	Poor	Tidak jelas dan tidak jernih
1	Bad	Sangat tidak jelas dan sangat tidak jernih

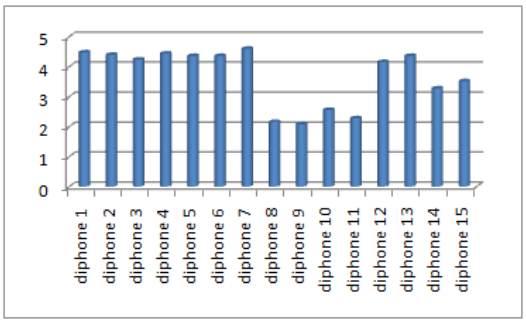
Perhitungan MOS :

$$\text{MOS} = \frac{\sum_{i=1}^{n} x(i).k}{N}$$

Dimana :

X(i) = Nilai sample ke i
K = Jumlah bobot
N = Jumlah pengamatan

Dari nilai perhitungan MOS untuk masing-masing diphone didapatkan bentuk grafik MOS pengujian kualitas suara diphone.



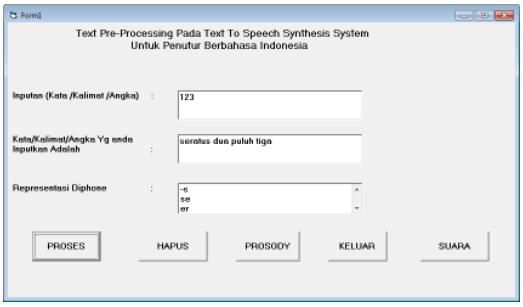
Gambar 5. Grafik MOS pengujian kualitas suara diphone

Dari hasil survey yang merupakan penilaian subjektif dari 25 responden dapat dikatakan bahwa diphone yang terdiri dari gabungan fonem vokal dan konsonan memiliki kualitas suara bagus atau suara yang jernih dengan nilai MOS rata-rata 4.41. Sedangkan diphone yang terdiri dari kombinasi fonem konsonan dengan konsonan memiliki kualitas suara yang tidak jelas dan tidak jernih dengan nilai MOS rata-rata 2.27. Kualitas suara yang tidak jelas lebih banyak diakibatkan oleh kesalahan pelabelan (penentuan batas-batas diphone) dalam pembuatan diphone database. Selain kesalahan proses pelabelan, kualitas suara diphone yang tidak jelas diakibatkan oleh perekaman kata-kata yang mengandung diphone bercampur dengan noise. Baik noise pada lingkungan sekitar ruang perekaman, maupun noise pada alat rekam kata yang digunakan.

4.2 Konversi Masukan Ke Representasi Diphone

4.2.1 Konversi Masukan Berupa Angka

Jika masukan sistem berupa angka, maka sistem akan mengkonversikan dari angka (numerik) ke string. Dari bentuk string inilah kemudian dikonversikan ke dalam bentuk representasi diphone. Angka yang dapat dikonversikan ke string hanya dibatasi sampai pada angka ratusan. Sebagaimana yang terlihat pada tampilan program dibawah ini.

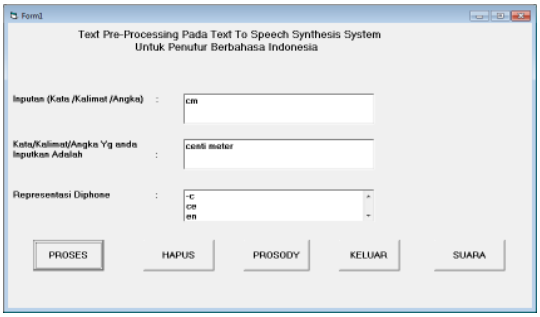


Gambar 6. Tampilan pengujian masukan angka

Dari tampilan program diatas, masukan sistem berupa angka yaitu "123". Selanjutnya masukan angka tersebut dikonversikan terlebih dahulu ke dalam bentuk string menjadi "seratus dua puluh tiga". Dari bentuk string inilah dilakukan pengonversian kedalam bentuk representasi diphone menjadi "-s, se, er, ra, at, tu, us, s- ,-d, du, ua, a-, -p, pu, ul, lu, uh, h-, -t, ti, ig, ga, a- ".

4.2.2 Konversi Masukan Berupa Akronim

Jika masukan sistem berupa akronim atau singkatan, pada program harus dilakukan inisialisasi terlebih dahulu untuk beberapa singkatan yang populer digunakan. Misalnya singkatan untuk "cm" maka kepanjangannya adalah "centi meter". Sebagaimana yang terlihat pada tampilan program dibawah ini.

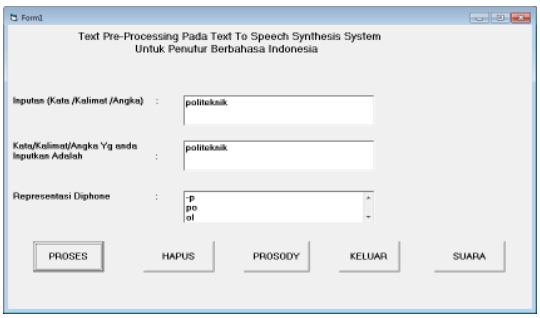


Gambar 7. Tampilan pengujian masukan akronim

Dari tampilan program diatas, masukan sistem berupa akronim atau singkatan yaitu "cm". Sebelumnya pada program, dilakukan inisialisasi untuk beberapa singkatan beserta kepanjangannya yang populer di sekitar kita. Misalnya "cm", "km", "kg", "jl". Untuk masukan "cm", maka sistem terlebih dahulu akan mengkonversikannya kepanjangan dari "cm" yaitu "centi meter". Dari kepanjangannya inilah dilakukan pengonversian ke bentuk representasi diphone menjadi " -c, ce, en, nt, ti, i-, -m, me, et, te, er, r- ".

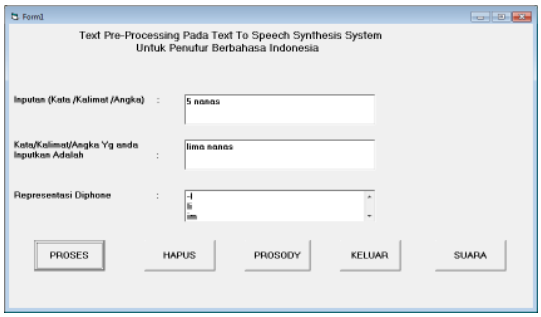
4.2.3 Konversi Masukan Berupa Kata / Kalimat

Jika masukan sistem berupa kata, maka program akan langsung mengkonversikan masukan ke dalam bentuk representasi diphone. Sebagaimana yang terlihat pada tampilan program dibawah ini.



Gambar 8. Tampilan pengujian masukan kata

Dari tampilan program diatas, masukan sistem berupa kata yaitu “politeknik”. Dari kata tersebut akan langsung dikonversikan ke dalam bentuk representasi diphone menjadi “ -p, po, ol, li, it, te, ek, kn, ni, ik, k- ”. Jika masukan sistem berupa kalimat yang terdiri dari angka dan kata, maka program akan mengkonversikan terlebih dahulu angka ke dalam bentuk string. Sebagaimana yang terlihat pada tampilan program dibawah ini.

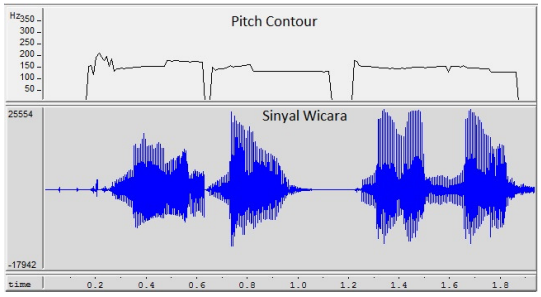


Gambar 9. Tampilan pengujian masukan kalimat

Dari tampilan program diatas, masukan sistem berupa kalimat yaitu “5 nanas”. Angka yang terdapat pada kalimat diatas, akan dikonversikan terlebih dahulu ke dalam bentuk string menjadi “lima nanas”. Selanjutnya kalimat tersebut dikonversikan ke dalam bentuk representasi diphone menjadi “ –l, li, im, ma, a-, -n, na, an, na, as, s- ”.

4.3 *Direct Concatenate* Antar Diphone

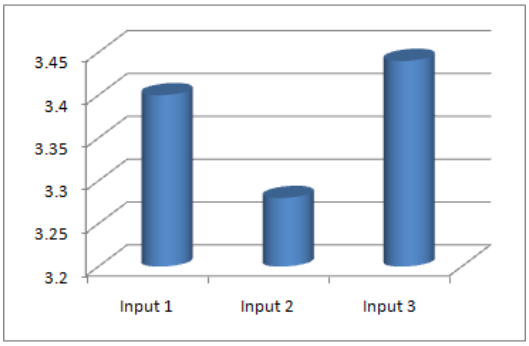
Pengujian pada *direct concatenate* (penyambungan langsung) juga dilakukan dengan melakukan survey ke beberapa responden terhadap kualitas suara pada *direct concatenate* antar diphone. Berikut ini adalah hasil pengeplotan pada masukan kalimat “5 nanas” beserta *pitch contour* sinyal wicara hasil penyambungan diphone tersebut.



Gambar 10. Hasil plot & *pitch contour* kalimat “5 nanas” dengan direct concatenate

Dari gambar diatas, terlihat *pitch contour* hasil penyambungan langsung antar diphone tidak rata. Hal ini menyebabkan suara hasil penyambungan tidak sama sehingga terdengar tidak begitu jelas.

Pada pengujian penyambungan langsung antar diphone *(direct concatenate)*, juga dilakukan dengan melakukan survey kepada beberapa responden terhadap kualitas suara penyambungan langsung antar diphone. Dan dari pengujian tersebut didapatkan nilai *Mean Opinion Score* (MOS). Responden pada pengujian ini sebanyak 25 orang yang dipilih secara acak. Dari perhitungan nilai MOS untuk masing-masing inputan didapatkan bentuk grafik MOS pengujian kualitas suara pada penyambungan langsung *(direct concatenate)*.



Gambar 11. Grafik MOS pengujian suara direct concatenate

Dari hasil survey yang merupakan penilaian subjektif dari 25 responden dapat dikatakan bahwa kualitas suara penyambungan langsung antar diphone cukup jelas dan cukup jernih dengan nilai MOS rata-rata 3.37. Jika dilihat pada hasil pengeplotan untuk masing-masing masukan diatas, terlihat *pitch contour* tidak rata antar penyambungan diphone. Hal ini yang mengakibatkan kualitas suara penyambungan tidak begitu jelas. Permasalahan ini dapat diatasi jika menggunakan algoritma *PSOLA (Pitch synchronous overlap and add)* pada proses *Prosody*. Sehingga dengan menerapkan algoritma PSOLA sebelum proses *concatenation* (penyambungan) diharapkan kualitas suara yang dihasilkan jauh lebih jelas jika menggunakan *direct concatenate* (penyambungan langsung).

5. KESIMPULAN

Setelah dilakukan pengujian dan analisa terhadap sistem, maka diperoleh kesimpulan sebagai berikut :

1. Pada pengujian masukan berupa angka, akronim, kata dan kalimat sistem telah dapat mengkonversikan ke bentuk representasi diphone dengan benar.
2. Kualitas suara pada diphone dipengaruhi oleh proses perekaman kata yang mengandung

diphone dan proses pelabelan (penentuan batas-batas diphone).

3. Kualitas suara pada penyambungan langsung antar diphone *(direct concatenate)* sebagaian besar tidak begitu jelas, namun suaranya masih dapat dikenali

6. DAFTAR PUSTAKA

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

[2] McLoughlin, Ian, *Applied Speech And Audio Processing,* Canbridge University Press, Singapore, 2009.

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

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

[5] Sudarwati, Rini., *Text To Speech Pada PC Dengan Metode Direct Generation (Pembuatan Generator Sinyal)*, PENS-ITS, Surabaya, 2005.

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

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

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

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