Jurnal Teknik Elektro CosPhi. Vol. 2. No. 2 Tahun 2018
ISSN (Print) : 2597-9329
ISSN (online) : 2597-9337

Text To Speech Bahasa Indonesia Menggunakan *Synthesizer Concatenation* Berbasis Fonem

Sudirman Melangi
Program Studi Teknik Informatika, Universitas Ichsan Gorontalo
Jln. Drs. Ahmad Najamuddin No. 17 Kota Gorontalo, 96138
*e-mail: oedhie.lidya@gmail.com

***Abstrak*— Penelitian tentang text-to-speech (TTS) telah dilakukan untuk berbagai bahasa dan untuk beberapa bahasa. Namun, beberapa masalah spesifik dalam pengembangan aplikasi TTS belum sepenuhnya terpecahkan. Semua pendekatan yang diusulkan perlu menciptakan sistem TTS yang memiliki kejelasan dan kealamian. Karena alat/platform yang tersedia saat ini untuk penerapan TTS memerlukan sistem yang dapat mengurangi penggunaan memori dan kesederhanaan dalam prosesnya. Dalam penelitian ini, penggabungan synthesizer digunakan dengan kombinasi pendekatan baru yang menggunakan basis data ujaran fonem dasar. Phonem adalah unit terkecil dalam sebuah ucapan. Penggunaan fonem diharapkan menghasilkan penggunaan memori yang rendah dan proses yang cepat. Selanjutnya menggunakan metode synthesizer concatenation dengan pendekatan algoritma Time Domain PSOLA (*Pitch Synchronous Overlap Add*).Metode ini diharapkan menghasilkan ucapan bahasa Indonesia yang alami. Berdasarkan pengujian dari 45 responden untuk kriteria penilaian intelligibility diperoleh \ penilaian fluidity dengan uji MOS = 3,66 dan naturalness dengan uji MOS = 3,57.**

Kata kunci: ***Text-to-Speech, Phoneme, NLP, Synthesizer***

Abstract*— Research on text-to-speech (TTS) has been conducted for various languages and for several languages, the results are very satisfying. However, some specific problems in developing TTS applications have not been fully resolved. All proposed approaches need to create a TTS system that has clarity and naturalness. Because the tools / platforms currently available for the application of TTS, require a system that can reduce memory usage and simplicity in the process. Synthesizer concatenation has been proven to produce satisfying results in various languages. In this study, combining synthesizers was used with a combination of new approaches that used the basic phoneme speech database. Phonem is the smallest unit in a speech. The use of phonemes can be expected to produce low memory usage and a fast process. Furthermore, the synthesizer concatenation with the Time Domain PSOLA algorithm approach (Pitch Synchronous Overlap Add) is expected to produce natural Indonesian speech. Based on the testing of 45 respondents for intelligibility assessment criteria, MOS = 3.66 test, fluidity assessment with MOS test = 3.66 and naturalness with MOS test = 3.57.

Keywords: ***Text-to-Speech, Phoneme, NLP, Synthesizer***

I. PENDAHULUAN

Ucapan merupakan tindakan memproduksi suara melalui variasi tekanan udara yang dipancarkan oleh sistem artikulatoris. Sementara, speech synthesis adalah suara yang diproduksi atau buatan manusia melalui mesin (komputer). Oleh sebab itu proses tersebut dapat digolongkan dalam tiga hal yang saling berkaitan, antara lain bagaimana mesin tersebut membaca text yang masuk, bagaimana mesin tersebut dapat mengeluarkan suara, serta unsur-unsur lain yang saling berkaitan agar dapat menjadikan mesin tersebut bisa membaca dan berbicara [1].

Berbagai upaya pengembangan untuk menghasilkan kualitas suara yang lebih alami dalam speech synthesis dengan mengaitkan antara teknologi komputer dan bahasa komputasi. Komputer sebagai salah satu pengolah data dapat menghasilkan speech synthesis untuk berbagai macam kegunaan praktis. Pengembangan kemampuan komputer mengubah teks menjadi suara tersebut selanjutnya dikenal sebagai *Teks-to-Speech System.*

Terdapat beberapa tantangan dalam penelitan di bidang sintesis ucapan yaitu kejelasan (*intelligibility*), kealamian (*naturalness*), efektifitas biaya (*cost-effectiveness*) dan ekspresifitas (*expressivity*). Di bidang industri tidak hanya mengharapkan kejelasan, kealamian, tetapi juga efektifitas biaya. Diharapkan sintesis ucapan hanya membutuhkan beberapa megabyte untuk penyimpanan (small footprint), kapasitas kecil CPU (*small CPU*), mudah dikembangkan ke bahasa lain dan kemungkinan menciptakan suara baru secepat mungkin.

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

II. STUDI PUSTAKA

Tujuan umum dari sebuah sistem *text-to-speech* (TTS) adalah untuk meniru gaya *pronunication* manusia dalam rangka untuk mengucapkan dengan jelas, alami, dan fasih untuk teks masukan terbatas. Sistem TTS pertama disajikan pada tahun 1968 untuk bahasa Inggris. Sejak saat

itu, banyak sistem TTS telah diusulkan untuk berbagai bahasa. Di masa lalu, sistem TTS biasanya mengadopsi pendekatan berbasis aturan untuk menghasilkan informasi prosodis [2]-[4].

Meskipun beberapa dari sistem TTS tersebut telah terbukti memiliki kinerja tinggi, akan tetapi masih tetap kesulitan untuk menyimpulkan secara manual satu set aturan yang tepat untuk sintesis berkualitas tinggi. Dalam beberapa tahun terakhir, telah diusulkan pendekatan baru yang menggunakan model statistik atau jaring saraf secara otomatis untuk mempelajari aturan-aturan dari satu set data yang besar. Hal ini biasanya disebut sebagai pendekatan data-driven. Efektivitasnya telah dikonfirmasi dengan banyak contoh yang sukses. Pada dasarnya, ini jauh lebih sederhana dibandingkan dengan pendekatan berbasis aturan konvensional karena kesulitan secara manual menganalisis aturan pengucapan manusia. Namun saat ini, kinerja tinggi sistem TTS telah menunjukkan peningkatan untuk banyak bahasa termasuk Inggris, Jepang, Jerman, China dan berbagai bahasa lainnya [5]

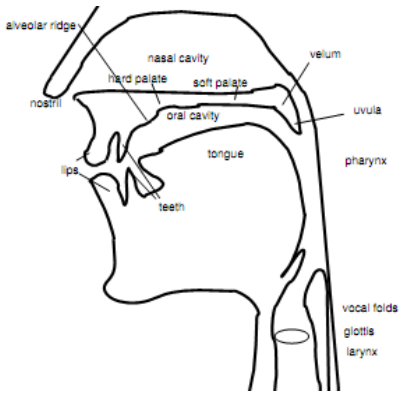
A. Suara

Suara adalah urutan gelombang tekanan yang merambat melalui media kompresibel (udara dan air). Gelombang tekanan tersebut dapat dipantulkan, dibiaskan ataupun dilemahkan oleh media. Suara juga dapat didefinisikan sebagai bunyi yang dikeluarkan dari mulut manusia. Suara dapat terjadi ketika berbicara, bernyanyi, tertawa ataupun menangis [6].

B. Sistem Produksi Suara Manusia

Proses produksi suara pada manusia dapat dibagi menjadi tiga buah proses fisiologis, yaitu: pembentukan aliran udara dari paru-paru, perubahan aliran udara dari paru-paru menjadi suara, baik *voiced*, maupun *unvoiced* yang dikenal dengan istilah *phonation*, dan artikulasi yaitu proses modulasi/pengaturan suara menjadi bunyi yang spesifik [6].

Organ tubuh yang terlibat pada proses produksi suara adalah: paru-paru, tenggorokan (*trachea*), laring (*larynx*), faring (*pharynx*), pita suara (*vocal cord*), rongga mulut (*oral cavity*), rongga hidung (*nasal cavity*), lidah (*tongue*), dan bibir (*lips*), seperti dapat dilihat pada gambar 1.



Gambar 1 Organ Pembentuk Suara manusia [6]

C. Transkrip Bunyi Bahasa

Transkripsi adalah suatu cara pengalihan penuturan yang berwujud bunyi ke dalam bentuk tulisan, kalimat, menggunakan lambang bunyi atau lambang fonetis (*phonetic symbol*). Tujuannya yaitu agar pengucapan bunyi fonem, morfem sesuai dengan ejaan yang berlaku pada suatu bahasa yang menjadi objeknya. Organisasi fonetik internasional yaitu *The International Phonetic Association (IPA)* telah menetapkan simbol fonetik internasional yang disebut *The International Phonetic Alphabets* (IPA), dimana merupakan kumpulan lambang-lambang yang didasarkan pada alpabet latin dan diciptakan untuk keperluan memberikan semua bunyi bahasa yang ada didunia.

D. Fonetik

Fonetik merupakan bidang kajian ilmu pengetahuan yang menelaah bagaimana manusia menghasilkan bunyi-bunyian bahasa dalam ujaran, menelaah gelombang-gelombang bunyi bahasa yang dikeluarkan, dan bagaimana alat pendengaran manusia menerima bunyi-bunyi bahasa untuk dianalisis otak manusia.

E. Fonem dan Jenisnya

Pengertian fonem yaitu merupakan kumpulan dari jenis Bahasa yang memiliki bunyi ucapan gunanya adalah untuk memberikan perbedaan arti kata tersebut (*distingtif*). Sebuah fonem dapat bervariasi tergantung pengaruh pada lingkungan yang ditempati hal ini sering disebut sebagai alofon. Terdapat banyak bahasa yang ada baik itu merupakan bahasa nasional maupun bahasa daerah. Khusus bangsa Indonesia dimana memiliki banyak Bahasa daerah pada setiap provinsi yang tentunya mempengaruhi Bahasa Indonesia.

Dijelaskan dalam kajian fonologi, bunyi atau fon ditulis di dalam kurung [ ], fonem tertulis di dalam tanda dua garis miring / /, dan huruf di antara kurung ( ). Fonem diftong terdiri dari /ay/, /aw/, /ey/ dan /oy/, di dalam ejaan dilambangkan dengan dua huruf vokal (ai), (au), (ei) dan (oi). Ada yang berpendapat bahwa diftong bukan fonem, sebab bunyi pertama adalah sebuah vokal dan bunyi kedua sebuah konsonan. Disebut diftong karena terpengaruh oleh sistem ejaan, yang seolah-olah menderetkan dua buah vokal dalam satu suku kata sebagai satu bunyi. Karena itu, dalam beberapa buku pelajaran, diftong sering diartikan sebagai “gabungan dua vokal yang diucapkan berurutan”. Ada yang berpendapat bahwa bunyi glotal bukan fonem sebab tidak ada pasangan minimal yang membuktikannya. Kalau ada pun cuma satu–satunya, yaitu antara [sakat] dan [saʔat]. Bunyi merupakan alofon dari fonem /k/, seperti muncul pada posisi akhir kata [bapa?], yang secara fonemis ditulis /bapak/ dan secara ortografis.

F. System Text to Speech Syinthesis

Pengertian secara umum sistem TTS sintesis merupakan sebuah aplikasi dengan bertujuan mengolah input berupa teks kemudian menghasilkan *output audio* yang menyerupai ucapan manusia [7]. Speech signal merupakan gelombang kontinu (gelombang akustik). Sebuah sinyal ucapan dapat dianalisis dengan berbagi metode pendekatan

karena mengandung nilai akustik, fonetik, fonologi, morfologi, sintatik, dan semantik

Adanya intonasi sebagai aspek akustik sinyal suara sangat membantu di dalam mengidentifikasi setiap segmen. Menurut metode formant dimana setiap fonem yang dihasilkan terutama oleh sistem vokal selama artikulasi mempengaruhi dinamika spektrum spektral suara. Variasi intonasi secara subtansial mempengaruhi idetitas pengucapan suatu kata sehingga fonem dapat menjadi panjang atau pendek, keras atau lemah, dan memiliki pola *pitch* (nada) yang bervariasi pula.

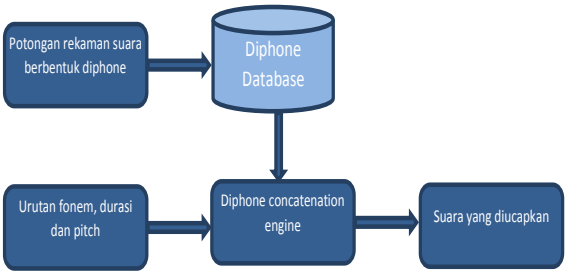
Secara umum sintesis ditentukan oleh 3 karakteristik yaitu:

1. Tingkat Pemahaman (intellegibility): Apakah teks yang dihasilkan komputer tersebut dapat dimengerti oleh pendengar (sejauh mana pendengar dapat memahami suara yang dihasilkan oleh komputer secara tepat).
2. Kelancaran Pengucapan (fluidity): Apakah teks yang dilafalkan oleh komputer tersebut lancer (transisi antara suku kata-kata lancar penyebutannya).
3. Kealamian Pengucapan (Naturalness): Apakah teks yang dilafalkan oleh komputer tersebut terdengar sesuai pengucapan manusia pada umumnya.

G. Concatenative Synthesizer

Umumnya metode ini memang menghasilkan output sintesa ucapan yang natural (mendekati ucapan manusia). Namun dari segi intelligibility (sejauh mana output ucapan tersebut dapat dimengerti oleh manusia) metode ini masih memiliki kekurangan. Proses metode ini dilakukan dengan cara menggabungkan (concatenate) unit-unit ucapan yang telah direkam sebelumnya.

Rekaman dari segmen-segmen ucapan tersebut dapat disimpan di dalam database dalam bentuk gelombang (uncoded) atau dalam bentuk encoded disesuaikan dengan pengkodingan ucapan (speed coding method) yang dipakai.



Gambar 2. Skema Proses Concatenative TTS

Gambar 2 menampilkan cara kerja sintesis concatenation berdasarkan pada rentetan bagian ucapan yang telah direkam. Dasar perangkaian sintesis adalah dengan menggabungkan segmen-segmen dari gelombang ucapan alami yang disimpan dalam database. Segmen-segmen tersebut dapat berupa kata-kata (*words*), unit sub-kata (*subword unit*) seperti fonem (*phonemes*), diphones dan suku kata (*syllables*). Sintesis ucapan perangkaian banyak digunakan luas, bekerja dengan prinsip pada pemilihan unit (*unit selection*). Sistem sintesis perangkaian pemilihan unit yang populer adalah *unisyn*, *clunits* dan *multisyn*.

H. Algoritma TD-Psola

Pada algoritma *Time Domain* PSOLA setiap perioda *pitch* akan dianalisis satu frame. Syarat yang dibutuhkan untuk ini adalah kemampuan mengidentifikasi permulaan perioda (*epochs*) pada *signal speech*. Gelombang ucapan dibagi menjadi frame-frame dengan menggunakan hamming window yang dikenakan pada perioda pitch sebelum dan sesudah epoch. bingkai hasil window ini kemudian dapat direkombinasikan dengan meletakkan epoch frame tersebut pada posisi yang sama seperti sebelumnya, kemudian menambahkan bagian yang bertumpang tindih (oleh sebab itu dinamakan overlap add). Setelah proses ini dilakukan, akan dihasilkan gelombang ucapan yang hampir mirip dengan gelombang ucapan yang asli.

Gelombang ucapan berupa frame-frame kecil dimana setiap *frame* tersebut dianggap sebagai *time invariant system*. Sebuah frame ucapan x[n] didapat dari gelombang penuh signal ucapan s[n] yang dikalikan dengan fungsi "jendela" (window) w[n] dalam domain waktu[6]:

$$x[n] = w[n]s[n] \qquad (2.1)$$

Proses windowing berfungsi untuk mengurangi efek diskontunuitas pada ujung-ujung frame yang dihasilkan oleh proses framing. Fungsi window yang digunakan dalam penelitian ini adalah *hamming window* yang memiliki persamaan sebagai berikut:

$$hamming\ \ w(n) = 0.54 - 0.46\ cos\left(\frac{2\times phi\times n}{(n-1)}\right), 0 \leq n \leq (n-1)$$

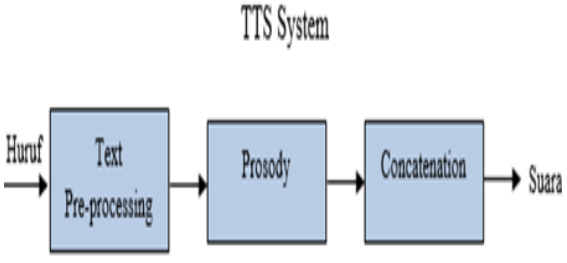
III. METODE

A. Text to Speech Synthesis System

Berikut direfresentasikan dalam bentuk diagram yang terdiri atas 3 bagian utama yaitu text pre-processing, pembangkitan prosody dan concatenation. Dibawah ini blok text to speech synthesis system:

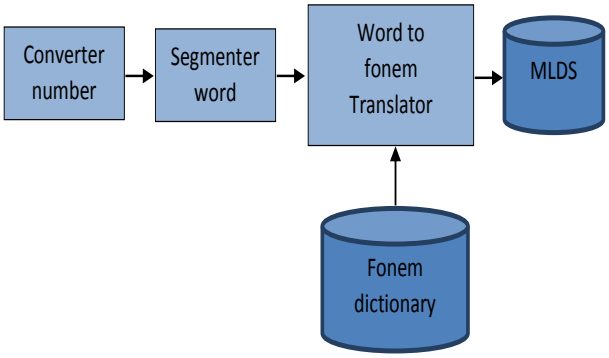
Metode berisi informasi tentang pelaksanaan penelitian, termasuk alur pelaksanaan penelitian, alat dan materi yang digunakan, tempat penelitian dan hal-hal lain yang dianggap perlu. Metode seharusnya ditulis secara rinci, dengan maksud agar pembaca yang berminat untuk mengulangi kembali penelitian ini, dapat melakukannya dengan informasi yang dituliskan pada bagian 'Metode'.

Posisikan gambar dan tabel pada bagian atas dan bawah kolom (jangan pada bagian tengah naskah); untuk menjamin kualitas gambar tetap bagus pada saat cetak, file-file gambar juga dikirimkan secara terpisah dengan format gambar dalam ekstensi EPS.



Gambar 3. *Bentuk umum Text to Speech Synthesis System*

Penggambaran untuk text pre-processing *sebagai berikut:*



Gambar 4. Alur Sistem *Text pre-processing*

Berdasarkan gambar 3 dan gambar 4 tersebut maka cara kerja sistem sebagai berikut:

- Konversi Angka
  sistem mengkonversikan angka ke dalam representasi fonem, jika masukan pada sistem berupa bilangan.
- Konversi Huruf/Kata
  sistem mengkonversikan kata atau kalimat ke dalam representasi diphone (gabungan dua buah fonem), jika masukan pada sistem berupa kata atau kalimat maka.
- Kumpulan Database Fonem
  Merupakan database yang berupa kumpulan dari fonem-fonem. Hasil penandaan pada sinyal ucapan.
- MLDS (*Multi Level Data Structure*)
  MLDS terdiri dari representasi fonem-fonem hasil pengkonversian inputan. *Database* ini diperlukan untuk proses pembentukan prosodi.

B. Pemilihan Teks Kalimat Bahasa Indonesia

Dari seleksi tersebut akan didapatkan minimum korpus kalimat bahasa Indonesia yang memenuhi keseimbangan fonetik (minimum phonetically balanced sentence corpus) yaitu kumpulan minimal kalimat-kalimat yang mengkover semua fonem yang ada didalam bahasa Indonesia (bahasa Indoneisa memiliki 26 huruf dan 32 fonem). Teks kalimat yang terpilih menjadi database (korpus) kalimat, yang selanjutnya akan dilakukan perekaman (recording) berdasarkan kalimat-kalimat tersebut.

C. Proses Perekaman Unit Ucapan

Dalam sistem TTS ini diperlukan basis data unit ucapan untuk menampung seluruh unit ucapan dalam hal ini yang dimaksudkan adalah fonem. Perbedaan bentuk fonem tergantung pada letak di dalam kata tersebut baik itu pada posisi awal suku kata, tengan ataupun akhir. Oleh karena itu setidaknya diperlukan 108 unit ucapan beberapa tambahan fonem khusus, perekaman unit ucapan disimpan dalam file raw atau wav.

D. Desain system

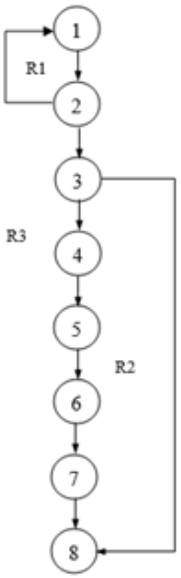
Pada tahap ini dirancang sebuah aplikasi yang akan menerima input berupa teks yang secara otomatis teks tersebut akan dikonversi oleh sistem menjadi fonem yang melibatkan prosodi menggunakan intonasi bahasa Indonesia. Selanjutnya fonem yang dihasilkan akan menjadi input untuk aplikasi yang menerapkan PSOLA. Adapun gambaran secara umum dari cara kerja aplikasi yang dirancang dibuat dalam dalam bentuk pemodelan beroreintasi objek seperti (use case diagram, sequence diagram, activity dan class diagram).

E. User Interface

Graphic User interface (GUI) dirancang untuk mempermudah user untuk berinteraksi dengan sistem dan melakukan pengujian pada metode yang digunakan. Untuk fase desain akan dilakukan dengan bantuan *tool* PHP untuk perancangan *user interface*.

F. Pengujian Sistem

Untuk mengetahui kinerja perangkat lunak secara internal maka harus dilakukan pengujian. Penelitian ini menggunakan teknik *White Box* untuk menguji alir program yang berisi proses pengucapan suara seperti gambar 5:



Gambar 5. Flowgraph Proses TTS

Dari flowgraph tersebut maka didapatkan:

Region (R) = 3
Node (N) = 8
Edge (E) = 9
Predikate Node = 2

Perhitungan:

1. V(G) = E – N + 2
   = 9 – 8 + 2
   = 3
2. V(G) = P + 1
   = 2 + 1
   = 3
3. CC = R1, R2, R3

Menentukan basis path:[27]

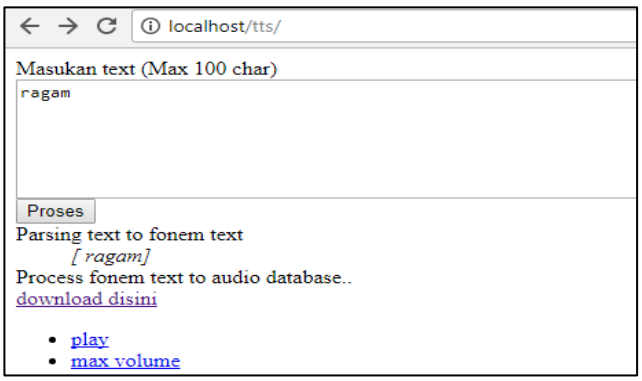
Basis 1 = 1, 2, 1, 2..
Basis 2 = 1, 2, 3, 8.
Basis 3 = 1, 2, 3, 4, 5, 6, 7, 8.

IV. HASIL DAN PEMBAHASAN

Berdasarkan hasil pembuatan perangkat lunak pada blok perancangan sistem, kemudian pengujian dilakukan dengan masukan kata, kalimat, atau angka ke dalam blok *text pre-processing*. Kata atau kalimat yang dimasukan akan dikonversikan kedalam bentuk representasi *fonem* dan *diphone* (gabungan dari 2 fonem). 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* seperti contoh di bawah ini.

Dari tampilan program diatas, masukan sistem berupa angka yaitu "1234". Selanjutnya masukan angka tersebut dikonversikan terlebih dahulu ke dalam bentuk string menjadi "seribu dua ratus tiga puluh empat". Dari bentuk string inilah dilakukan pengonversian kedalam bentuk representasi diphone menjadi "-s/er/ ri/bu/, -d/du/a-, ra/tu/us/s-,-t/ti/ig/ga/a-/, -p/pu/ul//lu/uh/h-/, -e/em/pa/t-".

Jika masukan sistem berupa kata atau kalimat, maka program langsung mengkonversikan masukan ke dalam bentuk representasi *fonem*. Contoh "ragam" yang disusun dari diphone-diphone menjadi "–r/ra/ga/am/m-".



Gambar 6. Proses *Text Pre-Processing* dengan Input Kata

Pada proses prosodi, dilakukan pemodelan terhadap representasi diphone hasil dari text preprocessing. Masing-masing diphone ditambahkan satu pitch periode data didepan dan dibelakang. Tujuan pemberian satu pitch periode data ini untuk mengaplikasikan algoritma PSOLA *(Pitch Synchronous Overlap Add Method)* untuk proses *concatenation*.

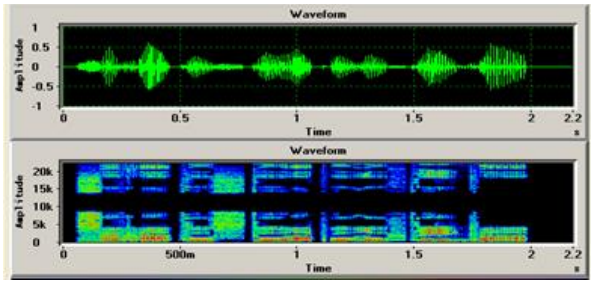
Panjang data dari *pitch* awal dan akhir sangat penting karena sangat menentukan berapa besar data yang di *overlap*-kan dan di tambahkan dengan *diphone* asli, hal itu yang akan menjadi dasar dalam penyambungan *diphone* dengan menggunakan metode PSOLA.

Pada metode PSOLA ada beberapa bagian yang menjadi prinsip dasar:

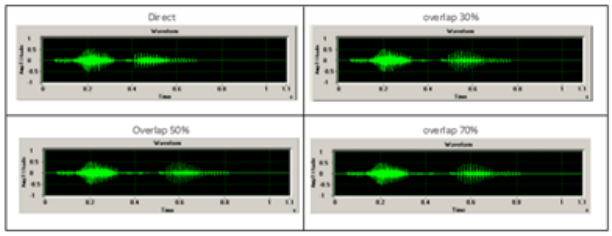
- Jika sinyal wav tersebut dalam proses *concatenation* terletak di paling depan, maka hanya satu *pitch periode* yang paling terakhir itulah yang akan di proses.
- Jika sinyal vaw tersebut dalam proses *concatenation* terletak di tengah, maka ujung yang paling depan dan yang paling akhirlah yang akan diproses.
- Jika sinyal wav tersebut dalam proses *concatenation* paling belakang, maka hanya satu *pitch periode* yang paling awal itulah yang akan di proses.

A. *Pengujian Kualitas Ucapan Sintesis*

Secara kualitatif sinyal ucapan sintesis dapat dipandang dalam representasi domain waktu dan spektogramnya. Dimana untuk memunculkan bentuk sinyal gelombang penulis menggunakan Bahasa pemograman Visual Basic 6.0.



Gambar 7. Bentuk Gelombang Ucapan Sintesis "1234"



Gambar 8. Bentuk Gelombang Ucapan Sintesis "Ragam"

Pada proses *concatenation* terjadi penggabungan kembali fonem-fonem pada proses prosody dengan menggunakan algoritma PSOLA *(Pitch Synchronous Overlap Add Method).* Pada PSOLA, proses *overlap* di lakukan setelah *fonem* asli di tambahkan satu *pitch* yang telah mengalami *windowing* di ujung depan dan belakang pada *fonem.* Selanjutnya panjang data dari masing-masing *pitch* kedua *fonem* yang akan disambungkan tersebut di kalikan dengan besar *overlap* yang akan di gunakan. Jika *overlap* 30%, maka panjang data dari masing-masing *pitch* dari kedua *fonem* yang akan disambungkan tersebut adalah 0.3 dari panjang data *pitch* tersebut. jika *overlap* 50%, dari masing-masing *pitch* dari kedua *fonem* yang akan disambungkan tersebut adalah 0.5 dari panjang data *pitch* tersebut, dan jika menggunakan *overlap* 70%, maka masing-masing *pitch* dari kedua *fonem* yang akan disambungkan tersebut adalah 0.7 dari panjang data *pitch* tersebut.

Dari hasil pengamatan, didapatkan bahwa overlap-add dari unit ucapan memberikan hasil yang masih dapat diterima, meskipun beberapa kali muncul 'klik/pop'. Secara umum bentuk sinyal ucapan di atas mirip, namun pada sinyal ucapan terlihat beberapa bagian sambungan yang masih kurang natural. Representasi spektogram dari kata "ragam" ditunjukkan dalam Gambar 8. Dari data yang ditampilkan di atas, terlihat kemiripan pada bentuk spektogram sinyal ucapan sintesis dan sinyal ucapan asli, dimana komponen-komponen frekuensi relatif sama.

B. *Pengujian Subjektif*

Pengujian yang terakhir dari sinyal ucapan hasil sintesis sistem TTS adalah pengujian MOS (Mean Opinion Score), dimana sejumlah pendengar langsung diperdengarkan sinyal ucapan sintesis dan memberikan penilaian dengan bobot/kriteria yang dijelaskan pada tabel 1.

Pengujian dilakukan kepada 45 orang pendengar, masing-masing pendengar menilai sepuluh macam kata ucapan yang berbeda. Dari data quisioner diperoleh hasil untuk kata "Ragam" yang disusun dari diphone-diphone menjadi "–r/ra/ga/am/m-" dan di nilai berdasarkan *Intelligibility* (tingkat pemahaman pendengar) sebagai berikut: Sangat Jelas (SJ) = 2, Jelas (J) = 27, Cukup Jelas (CJ) = 16, Kurang Jelas (KJ) = 0, Buruk (B) = 0. [5].

$$= (2 * 5) + (27 * 4) + (16 * 3) + (0 * 2) + (0 * 1)$$
$$= 10 + 108 + 48 + 0 + 0$$
$$= \frac{166}{45}$$
$$= 3{,}66$$

Jadi nilai MOS (*Mean Opinion Score*) *Intelligibility* = 3,66

Tabel 1. Kriteria Penilaian *Intelligibility* (Tingkat Pemahaman Pendengar Terhadap Ucapan)

No.	Kualitas	Kriteria
1	Bad	Ucapan tidak dapat dipahami, perangkaian ucapan sangat tidak jelas
2	Poor	Ucapan tidak dapat dipahami, perangkaian ucapan kurang jelas
3	Fair	Ucapan dapat dipahami, perangkaian ucapan cukup jelas
4	Good	Ucapan dapat dipahami, perangkaian ucapan jelas
5	Exellent	Ucapan dapat dipahami, perangkaian ucapan sangat jelas

Tabel 2. Kriteria Penilaian *Fluidity* (Tingkat Kelancaran Pengucapan)

No.	Kualitas	Kriteria
1	Bad	Ucapan tidak lancar, transisi antar suku kata sangat mengganggu
2	Poor	Ucapan tidak lancar, transisi antar suku kata mengganggu
3	Fair	Ucapan cukup lancar, transisi antar suku kata sedikit mengganggu
4	Good	Ucapan lancar, transisi antar suku kata nyaman
5	Exellent	Ucapan lancar, transisi antar suku kata sangat nyaman

Tabel 3. Kriteria Penilaian *Naturalness* (Tingkat Kealamian)

No.	Kualitas	Kriteria
1	Bad	Pengucapan datar (tidak berintonasi)
2	Poor	Pengucapan sedikit berintonasi, tidak sesuai pengucapan manusia pada umumnya
3	Fair	Pengucapan sedikit berintonasi, sesuai pengucapan manusia pada umumnya
4	Good	Pengucapan berintonasi baik, sesuai pengucapan manusia pada umumnya
5	excellent	Pengucapan identik dengan pengucapan manusia pada umumnya

Pengujian dilakukan kepada 45 orang pendengar, masing-masing pendengar menilai sepuluh macam kata ucapan yang berbeda. Dari data quisioner diperoleh hasil untuk kata "ragam" yang disusun dari diphone-diphone menjadi "–r/ra/ga/am/m-" dan di nilai berdasarkan *Fluidity* (tingkat kelancaran pengucapan) sebagai berikut: Sangat Jelas (SJ) = 1, Jelas (J) = 28, Cukup Jelas (CJ) = 16, Kurang Jelas (KJ) = 0, Buruk (B) = 0., Maka:

$$= (1 * 5) + (28 * 4) + (16 * 3) + (0 * 2) + (0 * 1)$$
$$= 5 + 112 + 48 + 0 + 0$$
$$= \frac{165}{45}$$
$$= 3{,}66$$

Jadi nilai MOS (*Mean Opinion Score*) *Fluidity* = 3,66

Selanjutnya pengujian dilakukan kepada 45 orang pendengar, masing-masing pendengar menilai sepuluh macam kata ucapan yang berbeda. Dari data quisioner diperoleh hasil untuk kata "ragam" yang disusun dari diphone-diphone menjadi "–r/ra/ga/am/m-" dan di nilai berdasarkan *Natural* (tingkat kealamian ucapan) sebagai berikut: Sangat Jelas (SJ) = 0, Jelas (J) = 26, Cukup Jelas (CJ) = 19, Kurang Jelas (KJ) = 0, Buruk (B) = 0, Maka:

$$= (0 * 5) + (26 * 4) + (19 * 3) + (0 * 2) + (0 * 1)$$
$$= 0 + 104 + 57 + 0 + 0$$
$$= 3{,}57$$

Berdasarkan data pengamatan subyektif, pendengar sudah dapat memahami rangkaian unit ucapan, dan pengucapan sistem *Text to Speech* cukup lancar. Namun demikian intonasi sinyal ucapan yang dihasilkan oleh sistem *Text to Speech* ini belum sama dengan ucapan asli dan masih belum natural. Untuk pengujian dengan kalimat yang sangat panjang, kualitas ucapan cenderung turun, baik dari segi kelancaran maupun tingkat pemahaman pendengar terhadap ucapan. Beberapa faktor dapat menjadi penyebab ini, antara lain kualitas *database* unit ucapan sangat menentukan kualitas sintesis. Penggunaan unit ucapan *fonem* memberikan keuntungan yang sangat signifikan dari penggunaan memori, akan tetapi oleh karena secara rata-rata durasinya sangat pendek, menyebabkan algoritma *sintesiser* PSOLA tidak dapat bekerja secara optimal.

V. KESIMPULAN

Berdasarkan proses pembangunan sistem yang telah dilakukan dalam penelitian ini maka dapat disimpulkan bahwa hasil secara umum dapat mengucapkan kata-kata dalam Bahasa Indonesia dengan cukup lancar dan dapat dimengerti oleh pendengar. Namun ucapan sintesis yang dihasilkan belum memiliki pola intonasi yang baik. Sintesiser perangkaian PSOLA dengan unit ucapan *database* fonem dapat bekerja dengan baik, meskipun masih terdapat *mismatch* pada perangkaian beberapa fonem, hal ini disebabkan oleh kualitas rekaman unit ucapan. Untuk menghasilkan kualitas pengucapan yang lebih natural, berintonasi dan mengandung emosi lagi penilis menyarankan untuk mengembangkan dengan penggunaan metode yang lain seperti sistem sintesis perangkaian ucapan statistik parametrik (HMN) dan sebagainya.

REFERENSI

[1] D. Thierry, "An Introduction To Text-To-Speech Synthesis," Netherlands, 1997.

[2] H. Mixdorff dan H. Fujisaki, "A Scheme for a Model-based Synthesis by Rule of F0 Contours of German Utterances," dalam *Proc. EUROSPEECH 95*, German, 1995.

[3] Y. Sagisaka, "On the Prediction of Global F0 Shape for Japanese Text-to-Speech," dalam *ICASSP*, 1990.

[4] Pressman, R.S. 2002. *Rekayasa Perangkat Lunak : Pendekatan Praktis (Buku I)*. Yogyakarta : Andi Yogyakarta.

[5] ITU-T, 2003: ITU-T Recommendation P.862.1: Mapping Functions for Transforming P. 862 Raw Result Scores to MOS-LQO.

[6] P. Taylor, "Text to speech synthesis," Cambridge , 2009.

[7] R. Kurzweil, "The Age Intelligent Machines," Cambridge, London, 1990.