SPEECH SYNTHESIZER BERBASIS DIPHONE MENGGUNAKAN ALGORITMA WAVEFORM SIMILARITY OVERLAP-ADD (WSOLA)

Adriyadi Subiyakto[1], Iwan Iwut Tritoasmoro[2], Gelar Budiman[3]

[1]Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom

Abstrak

Perkembangan speech synthesizer saat ini meningkat cukup pesat. Berawal dari hasil suara yang didapatkan tidak terdengar alami sama sekali, kemudian menuju ke arah prosodi yang semakin bagus. Salah satu contoh penerapan speech synthesizer adalah pada Text-to-Speech. Speech synthesizer berada pada blok terakhir dalam sistem Text-to-Speech. Speech synthesizer merupakan sebuah sistem yang mampu menghasilkan suara tiruan manusia dengan sintesa ucapan. Metode speech synthesizer yang ada saat ini adalah formant synthesis, articulatory synthesis, concatenative synthesis.

Metode yang digunakan dalam tugas akhir ini adalah metode diphone concatenation. Mula-mula sintesa ucapan dibentuk dengan melakukan perekaman suara dan hasilnya disimpan dalam database. Kemudian suara rekaman tersebut dipecah menjadi diphone yang memiliki transisi antar dua bunyi yang berdekatan (adjacent phones) sehingga akan lebih stabil saat digabungkan dengan diphone yang lain. Untuk menggabungkan unit ucapan diphone, digunakan algoritma Waveform Similarity Overlap-Add (WSOLA). Dengan menggunakan algoritma WSOLA, perangkaian antar diphone yang mengandung transisi antar dua bunyi yang berdekatan (adjacent phones), menjadi halus tanpa bunyi yang bersifat eksplosif.

Dari metode diphone concatenation dan penerapan algoritma WSOLA maka sintesis ucapan yang dihasilkan ternyata dapat dimengerti dengan jelas, lancar dalam pengucapan dan datar tanpa intonasi. Semakin beranekaragam unit diphone yang terdapat pada database akan memudahkan dalam pembentukan sintesis ucapan, sehingga akan meningkatkan kualitas hasil sintesis. Namun semakin besar memori yang dibutuhkan untuk menyimpan unit-unit diphone tersebut. Dari hasil penilaian Mean Opinion Score (MOS), parameter intelligibity mencapai nilai 3,41 dan fluidity yang mencapai 3,35 serta naturalness mencapai nilai 3,28. Dengan demikian kemampuan sistem dalam mensintesis suara ucapan manusia termasuk dalam kategori cukup.

Kata Kunci : speech synthesizer, diphone concatenation, algoritma WSOLA

Abstract

Currently, the development of speech synthesizer is increasing rapidly. From the speech results obtained that does not sound natural to hear at all, and then go to the better prosody. One of the speech synthesizer applications is on the Text-to-Speech system. Speech synthesizer is a system that is capable of producing sound imitation with human speech synthesis. There are three method of speech synthesizer, that is formant synthesis, articulatory synthesis, concatenative synthesis.

Diphone concatenation is the method used in this final assignment. Initially, speech synthesis formed with voice recording and the results stored in database. Then it brakes into diphone that has a transition between two neighboring sound (adjacent phones) so it will be more stable when combined with another diphone. To combine speech diphone units, Waveform Similarity Overlap-Add (WSOLA) algorithm is used. By using the WSOLA algorithm, combination between diphone containing transition between two neighboring sounds (adjacent phones), will be smooth without the explosive sound.

Combining diphone concatenation method with WSOLA algorithm would make the speech synthesis sounds clear, smooth and as natural as the voice of human being without prosody. The more diverse units of the diphone database will facilitate the establishment of speech synthesis, will improve the quality of synthesis results. But the larger memory will be needed to store them. Based on Mean Opinion Score (MOS), the intelligibility, fluidity and naturalness parameter has reach 3,41; 3,35 and 3,28. So that the system is fair enough in synthesizing human speech.

Keywords : speech synthesizer, diphone concatenation, WSOLA algorithm

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan *speech synthesizer* beberapa dekade terakhir meningkat cukup pesat. Berawal dari hasil suara yang didapatkan tidak terdengar alami sama sekali, hingga menuju ke arah sintesis ucapan yang semakin baik. Salah satu contoh penerapan *speech synthesizer* adalah pada *Text-to-Speech*. *Speech synthesizer* merupakan sebuah sistem yang mampu menghasilkan suara tiruan manusia dengan sintesis ucapan. Terdapat tiga metode dasar yang dapat digunakan pada *speech synthesizer* yaitu *formant synthesis, articulatory synthesis* dan *concatenative synthesis*. Tugas Akhir ini akan membahas penyusunan unit ucapan menggunakan *diphone concatenation synthesis*. *Diphone* dipilih sebagai unit ucapan karena *diphone* merupakan pelebaran titik tengah dari bagian *phone* yang stabil ke titik tengah *phone* berikutnya. Oleh karena itu, tiap *diphone* mengandung transisi antar dua bunyi berdekatan (*adjacent phones*) yang stabil sehingga dapat mengurangi distorsi saat perangkaian antar *diphone*. Selanjutnya, *concatenation synthesis* merupakan metode perangkaian segmen-segmen ucapan yang telah direkam sebelumnya. Metode *concatenation* dianggap sebagai salah satu metode yang sesuai untuk membuat sintesis ucapan berkualitas tinggi karena hasil keluaran sistem berupa sinyal suara sintesis yang terdengar alami dan dapat dimengerti dengan jelas.

Langkah awal dalam membentuk sintesis ucapan adalah dengan merekam berbagai macam kata dalam Bahasa Indonesia dan disimpan pada *database*. Kemudian, hasil rekaman tersebut disegmentasi menjadi unit-unit *diphone*. Perangkaian unit-unit *diphone* dilakukan dengan menggunakan algoritma *Waveform Similarity Overlap-Add* (WSOLA). Algoritma WSOLA memperkenankan menggunakan suara rekaman untuk dirangkai dengan lancar dan halus. Dengan menggunakan algoritma tersebut, perangkaian antar *diphone* yang mengandung transisi antar dua bunyi berdekatan diharapkan menjadi lebih halus dan dapat mengurangi bunyi yang bersifat letupan sehingga sintesis ucapan yang dihasilkan mampu memenuhi tiga kriteria, yaitu terdengar jelas dan dapat dimengerti, lancar dan alami (*intelligibility, fluidity, naturalness*).

1.2 Rumusan Masalah

Permasalahan yang diteliti dalam dalam tugas akhir ini adalah :

1) Bagaimana merancang suatu sistem *speech synthesizer* berbasis *diphone* dengan menggunakan metode *concatenation synthesizer*.
2) Bagaimana menerapkan algoritma *Waveform Similarity Overlap-Add* (WSOLA) sebagai teknik untuk menghaluskan perangkaian *diphone*.
3) Bagaimana menggunakan *database diphone* yang lengkap agar dapat melakukan perangkaian *diphone* dengan optimal.

1.3 Batasan Masalah

Beberapa permasalahan yang dibatasi adalah :

1) Metode *speech synthesizer* yang digunakan adalah *concatenation synthesizer* dengan unit ucapan *diphone*.
2) Algoritma yang diterapkan adalah *Waveform Similarity Overlap-Add* (WSOLA).
3) Format *diphone* sudah ditentukan dalam bentuk *.wav.
4) Kata yang diucapkan dalam Bahasa Indonesia dan mengacu pada Kamus Besar Bahasa Indonesia.
5) Persajakan (prosodi) tidak diperhatikan, sehingga suara yang keluar akan terdengar tanpa intonasi/datar.
6) Sistem hanya ditujukan untuk mengucapkan satu kata dalam Bahasa Indonesia.
7) *Database diphone* yang digunakan dalam aplikasi ini telah dibuat oleh Aggie Y Prihandi.

1.4 Tujuan dan Manfaat Penelitian

Adapun tujuan penyusunan tugas akhir ini adalah :

1) Meneliti dan merancang sistem *speech synthesizer* berbasis *diphone* dengan menggunakan metode *concatenation*.
2) Mempelajari dan menganalisis performansi yang dihasilkan oleh penerapan algoritma WSOLA yang digunakan untuk perangkaian *diphone.*

3) Menggunakan *database diphone* yang lengkap sebagai pendukung sistem agar didapatkan sintesis suara yang optimal.

Manfaat yang diharapkan pada penyusunan tugas akhir ini antara lain :

1) Dapat membuat suatu sistem *speech synthesizer* sederhana dengan menggunakan algoritma WSOLA.
2) Dapat menjadikan transisi antara dua *diphone* yang berdekatan menjadi lancar dan meminimalkan bunyi yang sifatnya eksplosif pada daerah perangkaian.
3) Mendapatkan hasil sintesis ucapan yang optimal, alami dan dapat dimengerti dari metode ini melalui simulasi dan pengujian yang dilakukan.

1.5 Metodologi Penelitian

Langkah – langkah yang digunakan dalam pengerjaan Tugas Akhir ini adalah :

1) Studi literatur

   Langkah ini dilaksanakan dalam bentuk :

   a. Mempelajari karakter-karakter *speech* seperti *pitch*, frekuensi *formant* dan energi
   b. Mempelajari metode *diphone concatenation synthesizer*
   c. Mempelajari konsep algoritma WSOLA

2) Perancangan Sistem

   Perancangan sistem berdasarkan algoritma yang telah dipelajari dan menyesuaikan dengan bahasa pemrograman yang digunakan.

3) Pengujian dan analisis

   Langkah ini terdiri dari :

   a. Menguji kemampuan algoritma WSOLA untuk menggabungkan unit-unit *diphone* dari hasil rekaman dengan menggunakan software MATLAB R2008a
   b. Menganalisis dan menyimpulkan hasil sintesis ucapan yang terdengar berdasarkan perangkaian unit-unit *diphone* yang dilakukan dengan algoritma WSOLA
   c. Penyusunan laporan tugas akhir dan kesimpulan akhir

1.6 Hipotesis

Pengerjaan tugas akhir ini diawali dengan menyusun hipotesis sebagai berikut :

1) Sintesis ucapan berbasis *diphone* dengan algoritma WSOLA akan memberikan hasil yang lancar, terdengar alami dan dapat dimengerti.
2) Semakin beranekaragam unit-unit *diphone* yang terdapat pada *database* akan memudahkan dalam pembentukan sintesis ucapan, namun semakin besar memori yang dibutuhkan untuk menyimpan unit-unit *diphone* tersebut.

1.7 Sistematika Penulisan

Tugas akhir ini disusun dalam lima bab, yaitu :

I. BAB I : Pendahuluan

Berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian, metodologi penulisan dan sistematika penulisan.

II. BAB II : Dasar Teori

Berisi tentang teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu tentang teori dasar *speech, pitch, formant, speech synthesis, concatenation synthesis* dan algoritma *Waveform Similarity Overlap-Add* (WSOLA).

III. BAB III : Perancangan Sistem

Berisi perancangan *concatenation synthesizer* dimana unit-unit *diphone* hasil rekaman suara yang ada pada *database* digabungkan dengan menggunakan algoritma WSOLA.

IV. BAB IV : Analisis Hasil Pengujian

Berisi analisis dari hasil simulasi mengenai sintesis ucapan berbasis *diphone* dengan menggunakan algoritma WSOLA.

V. BAB V : Kesimpulan dan Saran

Berisi kesimpulan dari analisis yang dilakukan dan saran untuk pengembangan lebih lanjut.

BAB V
KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Dari hasil analisa subjektif maupun objektif terhadap kinerja sistem, maka dapat ditarik kesimpulan sebagai berikut :

1. Penerapan algoritma WSOLA pada unit ucapan *diphone* berhasil menghasilkan sinyal sintesis ucapan yang jelas dan lancar.
2. Kualitas sinyal sintesis sangat bergantung pada kualitas unit ucapan yang disimpan dalam database.
3. Dengan menggunakan *diphone* sebagai unit ucapan dapat mengurangi terjadinya distorsi pada daerah penyambungan (*overlap*). Dan, semakin kompleks jenis unit ucapan yang digunakan, maka ukuran database akan semakin besar.
4. Pengolahan diphone dengan melakukan sinkronisasi *pitch* terlebih dahulu akan menghasilkan sintesis ucapan yang lebih datar atau tanpa intonasi jika dibandingkan dengan sinyal yang tidak melalui proses sinkronisasi pitch terlebih dahulu.
5. Algoritma WSOLA mampu mengucapkan kata dari *database* jenis kelamin yang berbeda tanpa mempengaruhi kerja sistem dalam mensintesis suara.
6. Berdasarkan hasil *Mean Opinion Score* (MOS), kemampuan sistem dalam mensintesis suara termasuk dalam kategori cukup, dengan perolehan nilai untuk parameter *intelligibility* sebesar 3,41; parameter *fluidity* sebesar 3,35 dan parameter *naturalness* sebesar 3,28.
7. Berdasarkan hasil uji *Mean Opinion Score* (MOS) untuk variasi nilai *overlap*, penggunaan nilai *overlap* yang akan menghasilkan sinyal sintesis yang optimal berada pada nilai 0,2 atau 20 %, dengan perolehan nilai untuk parameter *intelligibility* sebesar 3,90; parameter *fluidity* sebesar 3,86 dan parameter *naturalness* sebesar 3,76.

5.2 SARAN

1. Menambah kelengkapan *database* dengan membuat unit ucapan yang berkualitas tinggi.
2. Membuat *concatenation synthesizer* yang lebih baik dengan *triphone* sebagai unit ucapannya.
3. Membuat *concatenation synthesizer* dengan algoritma penyambungan yang lain, seperti *Linear Prediction* PSOLA (LP-PSOLA).

DAFTAR PUSTAKA

[1] **Arman, Ari Akhmad.** 2008. Konversi dari Teks Ke Ucapan. Departmen Teknik Elektro Institut Teknologi Bandung.

[2] **Arman, Ari Akhmad.** 2008. Proses Pembentukan dan Karakteristik Sinyal Ucapan. Departmen Teknik Elektro Institut Teknologi Bandung.

[3] **Arman, Ari Akhmad.** 2008. Teknologi Bahasa. Departmen Teknik Elektro Institut Teknologi Bandung. http://www.teknologibahasa.wordpress.com .diakses tanggal 4 Mei 2010

[4] **Fitriawati S., Atika.** 2009. Speech Synthesizer Berbasis Diphone Menggunakan Algoritma Time Domain Pitch Synchronous Overlap-Add. Tugas Akhir Fakultas Elektro dan Komunikasi IT Telkom : tidak diterbitkan

[5] **Prihandi, Aggie Y.** 2009. Speech Synthesizer Berbasis Diphone Menggunakan Algoritma Frequency Domain Pitch Synchronous Overlap-Add. Tugas Akhir Fakultas Elektro dan Komunikasi IT Telkom : tidak diterbitkan.

[6] **Verhelst W, Roelands M.** 1993. *An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech.* Belgium : Proceedings of the IEEE ICASSP-93

[7] ____. 2008. *Speech Synthesis.* [online]. *http://en.wikipedia.org/wiki/ Speech_synthesis*. diakses tanggal 4 Mei 2010.

[8] ____. 2008. *Concatenation.* [online]. *http://en.wikipedia.org/wiki/ Concatenation*. diakses tanggal 4 Mei 2010.