Uji Kemiripan Hasil Sintesis Suara Menggunakan Metode Jarak Mahalanobis

Yohanes Suyanto
Departemen Ilmu Komputer dan Elektronika FMIPA
Universitas Gadjah Mada
Yogyakarta
yanto@ugm.ac.id

Th. Prima Ari Setiyani
Jurusan Teknik Elektro Fakultas Sains danTeknologi
Universitas Sanata Dharma
Yogyakarta
ariprima@usd.ac.id

***Abstrak*—Telah dilakukan pengujian kemiripan sistem penerapan intonasi pada sintesis ucapan Bahasa Indonesia terhadap suara referensi menggunakan metode jarak Mahalanobis. Sistem ini akan mensintesis ucapan dengan pola intonasi yang diperoleh dari ekstraksi frekuensi dasar suara berita online. Teks untuk sintesis ucapan diperoleh dengan melakukan transkripsi suara berita online. Hasil sintesis dibandingkan dengan suara berita online asli menggunakan metode jarak Mahalanobis. Hasil penelitian menunjukkan bahwa nilai jarak Mahalanobis antara suara asli dan suara sintesis adalah rata-rata 0,94. Hasil ini lebih baik daripada suara sintesis dengan intonasi kaidah standar yaitu 1,20.**

Kata kunci—sintesis ucapan; intonasi; jarak Mahalanobis; Bahasa Indonesia

I. PENDAHULUAN

Sintesis ucapan adalah sebuah sistem berbasis komputer yang mampu membaca data teks menjadi suara ucapan. Sintesis ucapan terdiri atas bagian transkripsi, prosodi, dan sintesis. Transkripsi merupakan proses pengolahan teks menjadi satuan-satuan bunyi dan atribut yang menyertainya, sedang prosodi lebih menitikberatkan pada penentuan tinggi-rendah (*pitch*), panjang-pendek, dan keras-lemahnya suara. Bagian sintesis bertugas menyuarakan hasil pengolahan bagian transkripsi dan prosodi.

Untuk menuju pada suatu sistem sintesis ucapan Bahasa Indonesia yang secara otomatis menyertakan intonasinya maka dibuatlah penelitian sintesis ucapan dengan menyertakan pola-pola yang ada di beberapa contoh suara ucapan dalam Bahasa Indonesia.

Intonasi ucapan melibatkan tinggi rendah suara ucapan. Penerapan pola intonasi dalam sintesis ucapan diharapkan dapat menghasilkan sintesis ucapan yang cocok dengan intonasi pengucapan bahasa Indonesia.

Penelitian tentang pengujian kemiripan sintesis ucapan berdasar pola intonasi ini perlu dilakukan untuk lebih memfokuskan objek penelitian pada seberapa mirip hasil sintesis ucapan dengan suara aslinya.

Salah satu cara sintesis ucapan adalah dengan menggunakan metode penggandengan berdasar korpus suara yang ada [1]. Sintesis dilakukan dengan menggandeng potongan-potongan unit suara ucapan menjadi kata.

Dilain pihak menurut [2] diskontinuitas terdengar dalam sintesis ucapan berbasis penggandengan *concatenative*. Pengujian jarak diskontinuitas dan pengaruhnya terhadap pendengar dilakukannya dan menghasilkan bahwa model terbaik adalah dekomposisi AM & FM dekomposisi dari sinyal suara menggunakan diskriminan linier Fisher. Dengan cara ini masih terdapat unsur subjektivitas karena hasil didengar secara langsung oleh orang.

Proses memuluskan sambungan unit suara juga dilakukan oleh [3]. Dalam penelitiannya dia mengusulkan sebuah metode yang menggunakan skala spektrum setelah melakukan perataan spektrum dengan pendekatan sub-band linier untuk meminimalkan distorsi spektrum. Hasilnya dibandingkan dengan metode LPC (*Linear Predictive Coding*) dan *cepstrum*. Metode ini berusaha mencari sinyal dan distribusinya. Sinyal dinormalisasi sehingga distribusi terbesarnya dibawa ke nol. Hasil disajikan oleh tingkat distorsi spektrum untuk memperkirakan kinerja dari metode yang diusulkan. Tingkat distorsi spektrum berada di bawah rata-rata 2,12% dan menunjukkan bahwa metode yang diusulkan lebih baik dibandingkan dengan pendekatan yang ada lainnya.

Evaluasi terhadap hasil sintesis suara Bahasa Indonesia telah dilakukan oleh [4] dengan menggunakan metode PESQ. Hasilnya masih belum memuaskan karena dari nilai maksimum MOS (*Mean Opinion Score*) 4,5 hanya didapat rata-rata nilai MOS 1,2. Oleh karena itu perlu dilakukan penelitian dengan metode yang lain lagi. Pada penelitian kali ini akan digunakan metode jarak Mahalanobis.

II. METODE PENELITIAN

Penelitian dilaksanakan dengan mendasarkan pada basis data suara yang sudah ada yaitu dari sistem MBROLA [5]. Tahapan penelitian adalah:

1. Persiapan
2. Perekaman suara
3. Sintesis ucapan
4. Pembandingan nilai F0
5. Analisis hasil

A. *Persiapan*

Persiapan peralatan, bahan, dan referensi. Peralatan menggunakan komputer yang sudah tersedia sedang bahan berupa basisdata diambil di situs MBROLA. Referensi mengacu pada sintesis ucapan secara umum dan penentuan F0 atau frekuensi dasar.

Basis data bahasa Indonesia bernama id1 yang ini disediakan oleh proyek MBROLA (http://tcts.fpms.ac.be/synthesis). Fonem vokal yang tersedia adalah: a, e, e, i, o, dan u. Foenm diftong yang disediakan adalah: ai, oi, dan au. Kelompok fonem yang paling banyak yaitu konsonan yang meliputi: p, b, t, d, k, g, , c, j, f, s, z, h, m, n, ng, r, l, w, y, dan ny.

B. *Perekaman suara*

Perekaman suara sebagai model intonasi untuk diambil frekuensi dasarnya (F0) dengan menggunakan aplikasi Praat [6]. Hasilnya berupa vektor F0.

Praat menyediakan fungsi analisis suara (*speech analysis*), pelabelan dan segmentasi, algoritme pembelajaran, grafik, manipulasi suara, statistik, dan sintesis ucapan. Analisis suara dapat digunakan untuk mencari F0 dari sinyal suara dengan cara memilih menu Show analysis dan memberi cek pada Show Pitch, dan tidak mencek lainnya.

C. *Sintesis ucapan*

Sintesis ucapan dari teks dan intonasi yang sesuai dengan pola intonasi. Hasil sintesis juga diambil frekuensi dasarnya. Teks yang diumpan ke mesin sintesis ucapan didapat dari transkrip perekaman suara tahap sebelumnya. Sintak yang digunakan untuk menghasilkan sintesis ucapan adalah

```
mbrola id1 <berkas-pho> <berkas-wav>
```

`Mbrola` adalah nama aplikasi, `<berkas-pho>` adalah berkas hasil konversi teks ke format pho dan `<berkas-wav>` adalah hasil aintesis suara.

D. *Pembandingan nilai F0*

Pembandingan nilai F0 hasil sintesis ucapan dengan suara asli menggunakan metode jarak Mahalanobis. Jarak mahalanobis merupakan jarak euclid yang distandardisasi atau yang digeneralisasi. Jarak ini dapat mengatasi masalah perbedaan skala dalam data. Jarak mahalanobis dihitung berdasarkan persamaan:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}},$$

dengan s_i adalah deviasi standar dari x_i dan y_i, sedangkan x dan y adalah vektor sampel. Dalam penelitian ini x adalah F0 dari suara rekaman sedang y adalah F0 dari suara sintetis.

E. *Analisis hasil*

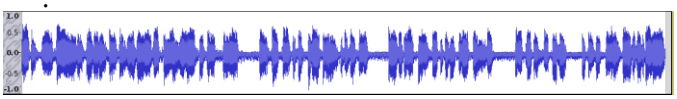
Analisis dilakukan terhadap hasil jarak mahalonobis untuk suara rekaman dan suara sintetis. Banyaknya suara yang diolah ada 5 pasang berkas.

III. HASIL DAN PEMBAHASAN

Penelitian dilakukan dengan mengambil rekaman suara dari seorang pembaca berita RRI. Dari rekaman tersebut diambil nilai frekuensi dasarnya. Selain itu dibuat juga transkripsi dari ucapan penyiar yang hasilnya berupa teks. Modul sintesis akan menyuarakan teks tersebut sebagai hasil sintesis suara. Suara inipun juga diambil nilai frekuensi dasarnya. Dari kedua nilai frekuensi dasar inilah diukur jarak Mahalanobisnya.

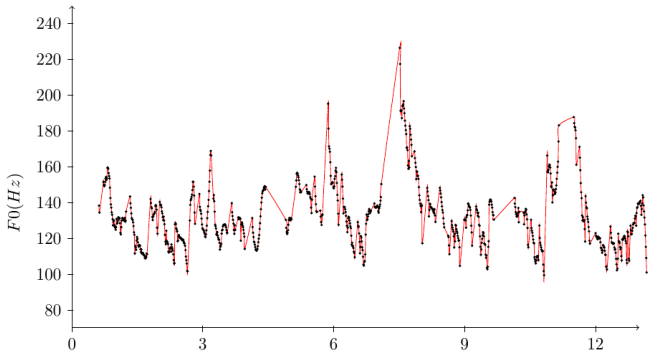
A. *Rekaman suara*

Berkas yang didapat dari pembaca berita berupa berkas suara .wav. Salah satu berkas yang diberi nama `kal101.wav` berbentuk gelombang suara terlihat pada Gambar 1.



Gambar 1. Visualisasi berkas suara kal101.wav

Rekaman tersebut kemudian diambil nilai F0 -nya dengan menggunakan aplikasi Praat Speech Analyzer. Hasilnya terlihat pada Gambar 2



Gambar 2. Grafik F0 untuk kalimat 1 (rekaman)

B. *Transkripsi*

Berkas tersebut kemudian dilakukan transkripsi di dapat teks

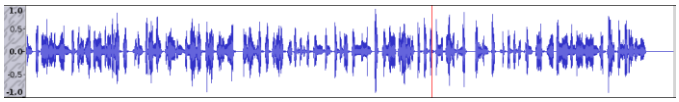
> *empat el es em lembaga swadaya masarakat di antaranya icewe indonesia korapsien wat pe es ha ka pusat studi hukum dan kebijakan menyampaikan aspirasi kepada panitia ed hok pah empat depede di kompleks parlemen senayan jakarta hari ini*

C. *Sintesis suara*

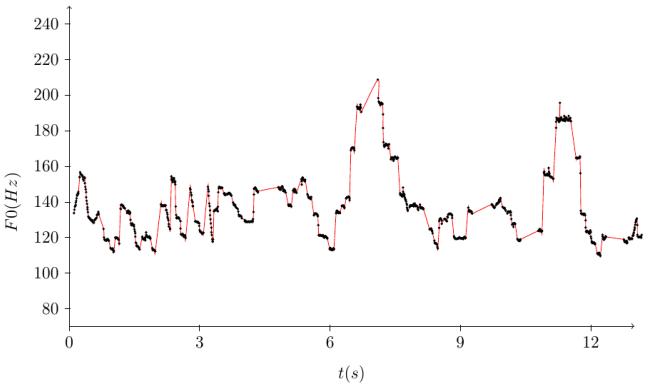
Selanjutnya teks tersebut digunakan sebagai masukan aplikasi sintesis ucapan, yang dalam hal ini menggunakan aplikasi Espeak. Hasil dari sintesis ucapan untuk teks tersebut terlihat pada Gambar 3. Suara tersebut kemudian juga diambil nilai F0-nya dengan menggunakan aplikasi Praat Speech Analyzer. Hasilnya terlihat pada Gambar 4

D. *Jarak Mahalanobis*

Jarak mahalanobis digunakan untuk menguji kemiripan 2 kelompok data yang mempunyai kolom (dimensi) sama tetapi jumlah data (baris) tidak harus sama.



Gambar 3. Visualisasi berkas suara kal101-sin.wav



Gambar 4. Grafik F0 untuk kalimat 1 (sintesis)

IV. PEMBAHASAN

Kalimat-kalimat hasil transkripsi yang digunakana dalam penelitian ini disimpan dalam berkas *antipreman, audax, dalambuku, ribuan,* dan *upacara.*

Berkas *antipreman* berisi '*Aksi pemasangan 200 spanduk di 80 titik strategis di kota Yogyakarta tersebut muncul secara spontan dari para pemuda dengan tujuan Yogyakarta menjadi aman dan bebas dari segala bentuk premanisme*'.

Berkas *audax* berisi '*Beragam komentar berhasil kami himpun usai seluruh peserta terakhir dengan jarak 100 kilometer memasuki garis finis*' sedangkan berkas *dalambuku* berisi '*Dalam buku tersebut tercantum sejarah perjalanan hotel dari masa ke masa.*'

Untuk kalimat *'Ribuan masyarakat memenuhi ruas jalan di depan pura Jagatnata yang terletakdi kelurahan Banguntapa kecamatan Banguntapan kabupaten Bantul.'* didapat bahwa jarak mahalanobis antara suara asli dengan sintesis **intonasi datar** adalah 1,74, sedang untuk **intonasi dengan kaidah umum** didapat 0,77. Jarak mahalanobis menjadi lebih baik untuk suara asli dan suara sintesis **perpaduan kaidah umum dan pola F0** yaitu mencapai 0,07.

Sintesis **intonasi datar** diperoleh dengan mengubah teks hasil transkripsi menjadi berkas pho dengan menerapkan frekuensi yang sama untuk setiap fonem yaitu sebesar 113 Hz. Penggalan hasil untuk kalimat '*Ribuan ...*' adalah

```
r   90 0 113
I   90 0 113
b   90 0 113
U   90 0 113
V   90 0 113
n   90 0 113
```

Sintesis **intonasi dengan kaidah umum** atau **kaidah standar** diperoleh dengan menerapkan pola frekuensi sesuai dengan pola yang didapat pada aplikasi espeak. Penggalan hasil pho untuk kalimat '*Ribuan ...'* adalah

```
r   78
I   42   0 94 20 95 40 96 59 97 80 99 100 99
b   78
U   59   0 117 80 109 100 109
V   61   0 110 80 106 100 106
n   79  100 98
```

Sintesis yang terakhir yaitu **perpaduan kaidah umum dan pola F0** diperoleh dengan menerapkan kaidah umum kemudian direvisi dengan pola F0 dari basis data yang sudah ada. Penggalan hasil berkas pho dengan metode sintesis ini untuk kalimat *'Ribuan ...'* adalah

```
r   78
I   42  0 107 20 108 40 109 59 110 80 113 100 113
b   78
U   59   0 116 80 108 100 108
V   61   0 122 80 118 100 118
n   79   100 117
```

Kalimat *'Upacara tradisi mubeng desa pengerupukan juga di gelar di pura Widyadarma Wedomartani Sleman.'* menghasilkan nilai untuk intonasi datar, standar, dan sintesis berturut-turut 1,12; 0,23; dan 1,02.

Dengan demikian hasil perpaduan sintesis dengan intonasi kaidah umum dan pola F0 dapat mendekatkan kemiripan sintesis ucapan dengan suara asli, menurut uji kemiripan jarak mahalanobis. Hasil lainnya tercantum pada Tabel 1. Hasil rata-rata pada intonasi datar, standar, dan sintesis adalah 1,03; 1,20; dan 0,94.

Dengan hasil ini terlihat bahwa hasil sintesis dengan parameter intonasi kaidah dan pola F0 lebih baik atau lebih mirip aslinya daripada sintesis dengan parameter intonasi kaidah umum saja.

TABEL I. HASIL UJI KEMIRIPAN DENGAN METODE JARAK MAHALANOBIS

Nama berkas	datar	standar	sintesis
antipreman	1,22	1,09	0,89
audax	0,88	0.87	0.71
dalambuku	1,18	2,08	2,00
ribuan	0,77	1,74	0,07
upacara	1,12	0,23	1,02
Rata-rata	**1,03**	**1,20**	**0,94**

V. KESIMPULAN

Uji kemiripan dengan jarak mahalonobis frekuensi fundamentalnya (F0) menunjukkan bahwa hasil akhir sintesis lebih mendekati suara asli dengan nilai 0,94 dibandingkan dengan sintesis dengan intonasi dengan kaidah standar yaitu 1,20. Perlu dilakukan ujicoba kemiripan dengan mengatur lebar jendela penghitungan frekuensi fundamental.

REFERENSI

[1] T. Dutoit, "Corpus-Based Speech Synthesis," In J. Benesty, M. M. Sondhi, and Y. A. Huang, editors, *Springer Handbook of Speech Processing*, pp. 437–456, Berlin Heidelberg: Springer, 2008, 10.1007/978-3-540-49127-9 21.

[2] Y. Pantazis, and Y. Stylianou, On the Detection of Discontinuities in Concatenative Speech Synthesis," In *Progress in Nonlinear Speech Processing*, volume 4391 *of Lecture Notes in Computer Science*, pp. 89–100, Berlin / Heidelberg: Springer, 2007. 10.1007/978-3-540-71505-4 6.

[3] J. Kim, H. Hahn, U.-J. Yoon, and M. Bae, *Wireless Personal Communications*, 50:435–446, 2009.10.1007/s11277-008-9615-x.

[4] Y. Suyanto, "Pengujian Kemiripan Hasil Sintesis Ucapan dengan Menggunakan Metode Perceptual Speech Quality Measurement (PSQM)," Technical report, MIPA UGM, 2012.

[5] T. MBROLA, The MBROLA Home Page, 2009. http://tcts.fpms.ac.be/synthesis/mbrola/.

[6] P. van Lieshout, P. PRAAT Short Tutorial. University of Toronto, Graduate Department of Speech-Language Pathology, Faculty of Medicine, Oral Dynamics Lab (ODL), 2003.