Pembuatan Perangkat Basis Data untuk Sintesis Ucapan (*Natural Speech Synthesis*) Berbahasa Indonesia Berbasis *Hidden Markov Model* (HMM)

Elok Anggrayni, Sekartedjo, dan Dhany Arifianto
Jurusan Teknik Fisika, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111
e-mail: sekar@ep.its.ac.id, dhany@ep.its.ac.id

Abstrak—**Salah satu teknik sintesis ucapan adalah sistem statistik parametrik sintesis ucapan menggunakan Hidden Markov Model (HMM). HMM-*based speech synthesis system* (HTS) adalah *toolkit* terbuka (*open source*) yang dapat dengan mudah diperluas ke berbagai macam bahasa. *Speech synthesis* dalam bahasa Indonesia dengan menggunakan HTS masih belum pernah dikembangkan (*under-resourced*). Penelitian ini diawali dengan pembuatan basis data suara bahasa Indonesia melalui proses perekaman, kemudian diikuti dengan proses segmentasi simbol fonetik, dan pemberian label. Dalam penelitian ini diperoleh basis data dalam bahasa Indonesia sejumlah 1529 kalimat yang sesuai dengan kaidah keseimbangan fonetik (*phonetically balanced*), yaitu telah memenuhi 33 jenis fonem. Selain itu, diperoleh juga segmentasi dan labeling *dataset* sebanyak 100 kalimat hasil rekaman suara laki-laki dan 100 kalimat hasil rekaman suara wanita. Penyiapan perangkat lunak untuk menjalankan sistem sintesis ucapan berbahasa Inggris berbasis HMM telah dilakukan dengan mengaplikasikan HTS yang menggunakan *Festival framework* dan berhasil dengan baik. Berdasarkan hasil uji kualitas suara menggunakan uji subyektif, melibatkan 20 responden, diperoleh *naturalness* dengan nilai *Mean Opinion Score* (MOS) 3,4 untuk pengujian hasil *training speaker dependent* (SD) *training demo* dan 3,2 untuk pengujian hasil *speaker adaptation/adaptive* (SAD) *training demo*. Dengan demikian, *synthetic speech* yang dihasilkan dapat dikategorikan baik dan perangkat lunak yang dipakai dapat digunakan untuk melakukan perancangan sistem sintesis ucapan berbahasa Indonesia.**

Kata Kunci—**sintesis ucapan, Hidden Markov Model (HMM), HMM-*based speech synthesis (*HTS*)***

I. PENDAHULUAN

TEKNOLOGI dikembangkan untuk membuat alat atau sarana yang dapat membantu dan memberi kemudahan bagi manusia untuk melakukan kegiatan dalam hidupnya. Seiring dengan perkembangan teknologi, manusia selalu menginginkan peningkatan kualitas dan kepraktisan dari alat-alat tersebut. Oleh karena itu, dibentuklah mesin-mesin yang dapat berinteraksi dengan manusia. Teknologi ini disebut teknologi *human machine* [1]. Teknologi *human machine* bertujuan menciptakan mesin yang memiliki kemampuan mengartikan informasi yang diucapkan manusia, bertindak sesuai dengan informasi tersebut, dan berbicara untuk menyempurnakan pertukaran informasi. Dengan kata lain menciptakan suatu mesin dengan kecerdasan buatan sehingga dapat berinteraksi dengan manusia melalui suara. Penelitian ke arah tersebut masih tetap dilakukan untuk mendapatkan hasil yang maksimal.

Di Indonesia sudah mulai dikembangkan *speech recognition* dengan menggunakan prinsip *neural networks.* Prinsip *neural networks* didasarkan pada pengolahan suatu masukan dengan mengikuti suatu model, seperti Hidden Markove Model (HMM). Masukan tersebut diolah untuk menghasilkan keluaran yang diinginkan. Selain itu, proses *training* diperlukan untuk mengenali data masukan dengan cepat. Pada perkembangan teknik pengolahan suara, khususnya pada *speech recognition* dan *speech synthesis* terdapat peluang untuk membentuk interaksi alami antara mesin dengan manusia. Pada awalnya interaksi antara manusia dengan mesin dilakukan dengan menggunakan *keyboard*, akan tetapi terdapat suatu kasus dimana penggunaan *keyboard* tersebut tidak sesuai untuk interaksi pada keduanya. Pada awalnya masukan yang bisa diterima oleh komputer hanya berupa teks masukan saja, misalnya pada internet maupun telepon. Namun, mesin masih belum dapat mengenali masukan berupa suara. Oleh karena itu, perlu dikembangkan teknologi *speech processing* untuk masukan berupa suara.

Sampai saat ini *speech database* dalam bahasa Indonesia masih dalam tiga tipe, yaitu digit terisolasi, penghubung, dan kata-kata percakapan yang sangat sederhana. Oleh karena itu, masih diperlukan penelitian dan pengembangan basis data suara dalam bahasa Indonesia [2].

Prosedur penelitian diawali dengan pembuatan basis data suara bahasa Indonesia melalui proses perekaman untuk keperluan *training* dan *testing* algoritma komputasi dari program yang telah dibuat dengan basis Hidden Markove Model (HMM). Kalimat yang digunakan untuk membangun basis data suara bahasa Indonesia dibuat berdasarkan kaidah keseimbangan fonetik (*phonetically balanced*). Keseimbangan fonetik akan tercapai jika dalam basis data kalimat yang digunakan telah mencakup seluruh fonem yang terdapat dalam bahasa Indonesia. Jumlah fonem yang terdapat dalam bahasa Indonesia adalah sebanyak 33 fonem [3]. Selanjutnya dalam penelitian ini akan membangun sistem *natural speech synthesis* berbahasa Indonesia berbasis Hidden Markove Model (HMM).

II. URAIAN PENELITIAN

A. Proses Pengambilan Data

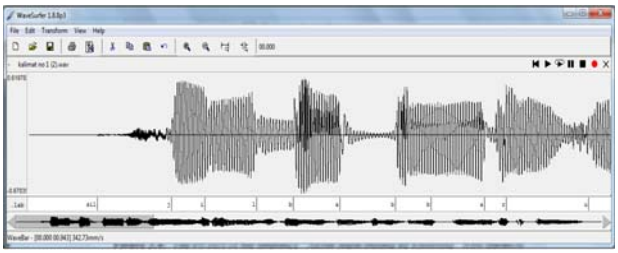
Pengambilan data dilakukan di ruang kedap suara Laboratorium Rekayasa Akustik dan Fisika Bangunan Jurusan Teknik Fisika ITS. Naracoba penelitian adalah rekaman suara satu orang laki-laki dan satu orang wanita. Naracoba diminta untuk duduk di depan mikrofon dengan jarak 2-3 cm. Dua orang yang direkam suaranya tersebut direkam pada saat yang berbeda dengan kalimat yang sama.

B. Langkah-Langkah Pembuatan Basis Data Suara

Basis suara yang digunakan dalam sintesis suara adalah hasil perekaman suara di ruang kedap suara dengan menggunakan kalimat bahasa Indonesia berdasarkan kaidah keseimbangan fonetik (*phonetically balanced)*. Basis data

Tabel 1.
Parameter Perekaman Basis Data Suara.

Parameter	Nilai
Frekuensi sampling	44100 Hz
Channel input/output	Mono
Bits/sample	32
Format file	wav



Gambar 1. *Software* wafesurfur untuk proses segmentasi dan *labeling*

ini terdiri atas suara laki-laki dan suara wanita sebanyak 1529 kalimat yang telah memenuhi 33 jenis fonem.

Perekaman menggunakan mikrofon *Shure* SM58 yang terhubung dengan E-MU 0404. Proses perekaman menggunakan pengaturan sebagai berikut:

a. Jarak antara mikrofon dengan naracoba ± 2-3 cm.
b. Non-aktifkan perangkat suara (*sound*) pada komputer sehingga dapat mendengarkan secara langsung suara yang direkam melalui E-MU 0404.
c. Parameter sinyal suara yang digunakan pada *software adobe audition* sesuai dengan yang parameter berikut ini dapat dilihat pada Tabel 1.

Contoh kalimat yang diucapkan naracoba adalah sebagai berikut:

/j-i-l-b-a-b b-a-r-u u-m-i b-e-r-w-a-r-n-a h-i-j-au m-u-d-a d-e-ng-a-n p-a-y-e-t b-u-ng-a d-i t-e-p-i-ny-a/

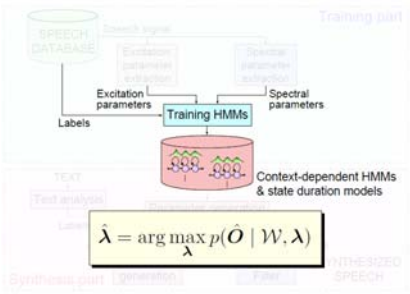
C. Segmentasi dan Labeling

Proses segmentasi dan *labeling* dilakukan dengan menggunakan *software* wavesurfur. Dari 1529 kalimat yang telah direkam diambil *sample* 100 kalimat dengan suara laki-laki dan 100 kalimat dengan suara wanita. 100 kalimat tersbut sama. Hasil segmentasi dan labeling dari 100 kalimat tersebut akan diolah secara statistik dan pemodelan untuk memprediksi tingkat kompresi energi bunyi untuk setiap orang. Model ini akan digunakan untuk estimasi tingkat *impairment* dan pada pita frekuensi berapa terjadi penurunan tingkat energi kompresi tersebut.

D. Ekstraksi Parameter F_0, Mel-Cepstrum, Delta Cepstrum, dan Delta-Delta Cepstrum

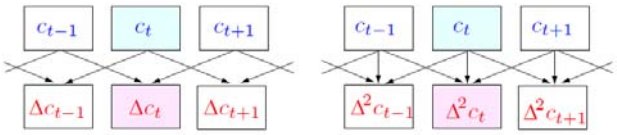
Pada HTS, vektor keluaran pada HMM terdiri dari dua bagian, yaitu spektrum dan eksitasi. Bagian spektrum terdiri dari koefisien melcepstral termasuk koefisien ke nol, delta, dan koefisien delta-delta. Selain itu, bagian eksitasi terdiri dari log frekuensi dasar (log F_0), delta, dan delta-delta. Dari hasil rekaman yang telah dilakukan, dapat dicari nilai parameter-parameter spektrum dan eksitasi tersebut.

Berikut adalah persamaan yang digunakan untuk mencari fitur dinamis pada HMM dapat dilihat pada Gambar 3. Urutan pengamatan frekuensi dasar (F_0) terdiri dari satu nilai dimensi yang kontinu dan simbol diskrit yang menggambarkan *"unvoiced"*. Oleh karena itu, HMMs diskrit atau kontinu tidak dapat diterapkan untuk pemodelan F_0. Untuk model urutan pengamatan tersebut digunakan jenis HMM-*based multi-space probability distribution* (MSD-HMM) [4]. MSD-HMM merupakan HMM diskrit

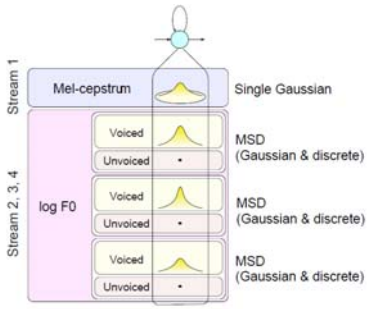


Gambar 2. Parameter eksitasi dan spektral [4]

$$\Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1})$$
$$\Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1}$$



Gambar 3. Fitur dinamis [4]



Gambar 4. Struktur distribusi *state-output* [4]

dan kontinu yang bercampur menjadi HMM sebagai permasalahan khusus. Melakukan pemodelan frekuensi dasar itu sulit karena terdapat sifat yang berbeda saat dilakukan pengamatan F_0 di daerah *voiced* dan *unvoiced*.

E. Langkah-langkah Sintesis Suara dengan Menggunakan HMM-based speech synthesis system (HTS)

1) Software

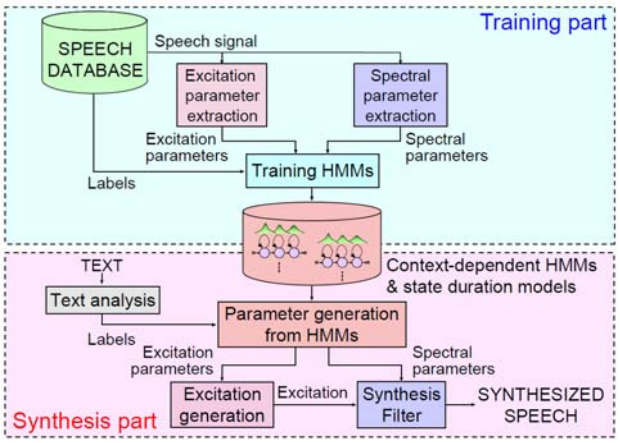
Sebelum melakukan *training* data dengan menggunakan HTS 2.2, harus ada beberapa program yang diinstall di komputer. Program tersebut antara lain:

a Festival (*a generic speech synthesis system*)
b SPTK-3.3
c HTS-2.2
d hts_engine API-1.03, dan
e OpenFst-.

III. PROSES TRAINING HMM-BASED SPEECH SYNTHESIS SYSTEM (HTS)

HMM-*based speech synthesis system* (HTS) dikembangkan oleh kelompok kerja HTS. Bagian *training* pada HTS telah digunakan sebagai versi modifikasi dari HTK dan dirilis sebagai bentuk kode *patch* untuk *Hidden Markov Toolkit* (HTK). Kode *patch* dirilis di bawah lisensi perangkat lunak bebas.

HTS pertama kali diimplementasikan untuk bahasa Jepang. HTS memiliki bahasa *dependent module* (daftar



Gambar 5. HMM-*based speech synthesis system* (HTS) [4]

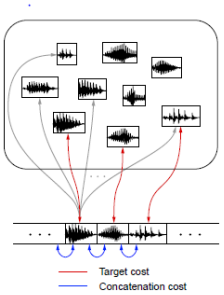
faktor-faktor kontestual). Dengan demikian, HTS dapat dengan mudah diperluas dan digunakan untuk berbagai macam bahasa. Keuntungan yang didapatkan dengan menggunakan HTS adalah mesin *runtime* yang dihasilkan berukuran kecil, yaitu kurang dari 1Mbytes, termasuk bagian analisis teks. Selain itu, HTS juga dapat mengubah karakter suara dari ucapan yang disintesis dengan menggunakan teknik *speaker adaptation* yang dikembangkan untuk pengenalan suara (*speech recognition*) [4]. Pada bagian *training*, parameter spektrum dan eksitasi diambil dari basis data suara dan dimodelkan dengan konteks *dependent* HMMs. Pada bagian sintesis, *context dependent* HMMs diubah sesuai dengan teks yang disintesis. Kemudian parameter spektrum dan eksitasi dihasilkan dari HMM dengan menggunakan parameter suara algoritma [8]. Modul *excitation generation* dan modul sintesis filter mensintesis gelombang suara menggunakan parameter eksitasi dan spektrum yang dihasilkan. Ekstraksi dari pendekatan dilakukan pada karakter suara dari ucapan yang disintesis sehingga dapat dengan mudah diubah dengan mengubah parameter-parameter HMM. Pada kenyataannya, proses ini menunjukkan bahwa kita dapat mengubah karakteristik suara dari ucapan yang disintesis dengan menerapkan teknik *speaker adaptation*, teknik interpolasi pembicara, dan teknik *eigenvoice*.

A. Implementasi HTS Implementation pada Arsitektur Festival

Pada HTS-demo_CMU-ARCTIC-SLT, script untuk mencari nilai ekstraksi F_0 sederhana script dituliskan Tcl/Tk. Script ini disebut get_f0, merupakan fungsi yang digunakan pada *open-source speech toolkit Snack*. Oleh karena itu, HTS-demo_CMU-ARCTIC-SLT juga harus dikompilasikan dengan Tcl/Tk with Snack.

Untuk proses *Setup* HTS-demo_CMU-ARCTIC-SLT dapat melakukan *running configure script* seperti di bawah ini:

```
% cd HTS-demo_CMU-ARCTIC-SLT
% ./configure --with-tcl-search-path=/usr/local/ActiveTcl/bin \
--with-fest-search-path=/usr/local/festival/examples \
--with-sptk-search-path=/usr/local/SPTK-3.3/bin \
--with-hts-search-path=/usr/local/HTS-2.1.1_for_HTK-3.4.1/bin \
--with-hts-engine-search-path=/usr/local/hts_engine_API-1.03/bin \
--with-openfst-search-path=/usr/local/openfst-1.1/bin
```



Gambar 6. *Unit Selection Scheme* [4]

Untuk memulai *running demonstration*, menggunakan script berikut:

```
% cd HTS-demo_CMU-ARCTIC-SLT
% make
```

Setelah menyusun data *training*, HMMs diperkirakan membutuhkan waktu sekitar 12 sampai 18 jam untuk melakukan sintesis.

IV. HASIL DAN PEMBAHASAN

A. Statistika Basis Data Kalimat

Untuk merancang sistem sintesis ucapan *(natural speech synthesis)* berbahasa Indonesia berbasis *Hidden Markov Model* (HMM) dibutuhkan basis data suara. Basis data suara yang dibuat berdasarkan kaidah keseimbangan fonetik (*phonetically balanced*). Dari basis data 1529 kalimat bahasa Indonesia yang telah dibuat, dilakukan proses statistika untuk menentukan berapa jumlah fonem konsonan, vokal tunggal, dan vokal rangkap agar dapat diketahui apakah basis data kalimat tersebut sudah memenuhi 33 jenis fonem.

B. Hasil Proses Training pada HMM-based speech synthesis system (HTS)

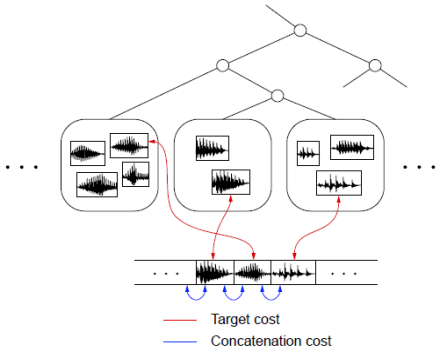
Perkembangan terbaru dalam sistem sintesis ucapan adalah statistik parametrik sistem sintesis ucapan berdasarkan Hidden Markov Model (HMM). *HMM-based speech synthesis system* (HTS) dikembangkan untuk mengatasi masalah pemilihan satuan sistem sintesis ucapan yang menyebabkan dari proses menjadi lambat saat melakukan sintesis karena penyimpanan yang sangat besar. HTS memiliki kemampuan untuk mensintesis ucapan dengan tingkat tinggi kealamian yang sebanding dengan unit sistem seleksi sintesis ucapan (*unit selection speech synthesis system*). Sistem ini pertama kali diusulkan oleh Yoshimura, Tokuda, dan Kobayashi, (1999).

Dengan menggunakan implementasi HTS pada arsitektur Festival, proses *training* data menggunakan 524 kalimat dari CMU *Communicator database*[3]. *Speech signal* disampel pada frekuensi 16 kHZ.

Faktor-faktor ini diambil dari ucapan-ucapan menggunakan ekstraksi fitur fungsi dari Festival *speech synthesis system*. Semua sistem di-*training* dalam beberapa jam, untuk *speaker dependent* membutuhkan waktu 18-20 jam dan untuk *speaker adaptation/adaptive* membutuhkan waktu sekitar 2 sampai 3 hari. Pohon resultan untuk model spektrum, model F_0, dan model *state duration* masing-masing totalnya bernilai 781, 1733, dan 1018 daun. Waktu *running* yang dibutuhkan mesin inti (*core engine*) terdiri dari 8 modul, *decision trees* untuk spektrum, F_0 dan durasi,

Tabel 2.
Ukuran Biner file pada HTS *Run-Time Engine*

module		size
decision tree	spectrum	102 kbyte
	F_0	156 kbyte
	duration	116 kbyte
distribution	spectrum	457 kbyte
	F_0	81 kbyte
	duration	39 kbyte
converter		3 kbyte
synthesizer		34 kbyte
total		988 kbyte



Gambar 7. Pengelompokkan Berdasarkan *Unit Selection Scheme* [5]

distribusi spektrum, F_0 dan durasi, sebuah konverter yang mengubah fitur yang telah diekstraksi oleh Festival ke dalam urutan label *context dependent* dan *synthesizer* yang menghasilkan gelombang untuk diberikan urutan label. Ukuran biner file dari masing-masing modul dapat dilihat pada Tabel 2.Dengan mendengarkan contoh *synthesized speech* pada:

`http://kt-lab.ics.nitech.ac.jp/~zen/sound/`

Sehingga dapat meyakinkan bahwa prosodi ini cukup alami. HTS juga dapat digunakan untuk *prosoy predictor* pada unit seleksi berbasis *speech synthesis system*.

C. Pembahasan

Gambar 8 menjelaskan mengenai lebar yang digunakan *unit selection scheme*. Pada *scheme* ini dapat didefinisikan bahwa jarak heuristis antara konteks untuk mengukur nilai target. Untuk menghindari hal ini, pengelompokan berdasarkan *scheme* juga dikembangkan [6].

Setelah menyusun data s*peaker dependent* (SD) *training demo*, HMMs membutuhkan waktu sekitar 24 jam untuk melakukan sintesis. Pada *training* ini dihasilkan 40 kalimat dengan karakter suara baru yang dihasilkan. Untuk menilai kualitas suara yang dihasilkan setelah proses *training* menggunakan HTS, dilakukan uji MOS *(Mean Opinion Score*) untuk melihat penilaian subyektif. MOS ini merupakan nilai rata- rata penilaian subyektif dari suatu obyek berdasarkan skala angka.

Berdasarkan hasil uji kualitas suara menggunakan uji subyektif, melibatkan 20 responden, diperoleh *naturalness* dengan nilai *Mean Opinion Score* (MOS) 3,4 (menunjukkan nilai cukup) untuk pengujian hasil *training speaker dependent* (SD) *training demo*. Dari pengujian hasil *training* s*peaker adaptation/adaptive* (SAD) *training demo*, *score* rata-rata kualitas yang diperoleh cukup tinggi. Hal ini dapat dilihat pada gambar 19. Rata-rata keseluruhan nilai untuk kelimabelas kalimat uji adalah 3,2 (menunjukkan nilai cukup). Rata-rata *score* diantara kelimabelas kalimat hampir sama.

Berdasarkan penyiapan perangkat lunak untuk menjalankan sistem sintesis ucapan berbahasa Inggris berbasis Hidden Markov Model (HMM) dengan mengaplikasikan HMM-*based speech synthesis* (HTS) yang menggunakan *Festival framework*, perangkat lunak berhasil dengan baik. Dapat dilihat juga berdasarkan hasil uji kualitas suara dengan menggunakan metode MOS ((*Mean Opinion Score*) untuk pengujian hasil *training speaker dependent* (SD) *training demo* dan s*peaker adaptation/adaptive* (SAD) *training demo* menunjukkan bahwa nilai kualitas suara adalah cukup. Dengan demikian, *synthetic speech* yang dihasilkan dapat dikategorikan baik dan perangkat lunak yang dipakai dapat digunakan untuk melakukan perancangan sistem sintesis ucapan (*natural speech synthesis*) berbahasa Indonesia.

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, didapatkan kesimpulan bahwa:

a. Diperoleh basis data dalam bahasa Indonesia untuk pembuatan sintesis ucapan *(natural speech synthesis)* berbahasa Indonesia berbasis Hidden Markov Model (HMM) sejumlah 1529 kalimat yang sesuai dengan kaidah keseimbangan fonetik (*phonetically balanced)*, yaitu telah memenuhi 33 jenis fonem.
b. Diperoleh segmentasi dan *labeling dataset* sebanyak 100 kalimat hasil rekaman suara laki-laki dan 100 kalimat hasil rekaman suara wanita untuk mempersiapkan rancangan sistem sintesis ucapan (*natural speech synthesis*) berbahasa Indonesia berbasis Hidden Markov Model (HMM).
c. Penyiapan perangkat lunak untuk menjalankan sistem sintesis ucapan berbahasa Inggris berbasis HMM telah dilakukan dengan mengaplikasikan HTS yang menggunakan *Festival framework* dan berhasil dengan baik. Berdasarkan hasil uji kualitas suara menggunakan uji subyektif, melibatkan 20 responden, diperoleh *naturalness* dengan nilai *Mean Opinion Score* (MOS) 3,4 untuk pengujian hasil *training speaker dependent* (SD) *training demo* dan 3,2 untuk pengujian hasil *speaker adaptation/adaptive* (SAD) *training demo*. Dengan demikian, *synthetic speech* yang dihasilkan dapat dikategorikan baik dan perangkat lunak yang dipakai dapat digunakan untuk melakukan perancangan sistem sintesis ucapan berbahasa Indonesia.

Untuk penelitian selanjutnya sebaiknya suara yang direkam ditambahkan dengan suara naracoba profesional, seperti penyiar radio, pembawa berita di radio atau televisi. Kemudian hasil sintesis suara antara naracoba biasa dengan naracoba profesional dibandingkan hasilnya.

UCAPAN TERIMA KASIH

Terima kasih kepada dosen pembimbing, seluruh dosen dan staff pengajar jurusan Teknik Fisika, dan seluruh Mahasiswa Teknik Fisika, atas bantuan dan dukungan yang diberikan sehingga makalah ini dapat diselesaikan sesuai dengan yang direncanakan.

DAFTAR PUSTAKA

[1] Tolba, hesham and Douglas O'Shaughnessy. "Speech Recognition by Intelligent Machines". IEEE Press, 2001.

[2] Lestari, Dessi Puji, Nonmember and Sadaoki FURUI, and Fellow, Honorary Member. IEICE TRANS. INF & SYST., VOL.E93-D, NO.9 SEPTEMBER 2010.

[3] Suyanto. 2007. "An Indonesian Phonetically Balanced Sentence Set for Collecting Speech Database". Jurnal Teknologi Industri Vol. XI No. 1 Januari 2007: 59-68.

[4] Tokuda, Keiichi and Heiga Zen. 2009. Fundamentals and Recent Advances in HMM-Based Speech Synthesis. Nagoya Institute of Technology: Toshiba Research Europe.

[5] A. W. Black and N. Campbell. "Optimising selection of units from speech databases for concatenative synthesis". Proc. EUROSPEECH, pp.581-584, Sep 1995.

[6] A. W. Black and P. Taylor. "Automatically clustering similar units for unit selection in speech synthesis". Proc. EUROSPEECH, pp.601–604, Sep 1997.

[7] Dey, Subhrakanti. "Reduced-Complexity Filtering for Partially Observed Nearly Completely Decomposable Markov Chains". IEEE Transactions on Signal Processing, Vol. 48, No. 12, Desember, 2000.

[8] Dugad, R., dan Desai UB. 1996. A Tutorial on Hidden Markov Models. India: Technical Report, Department of Electrical Engineering, Indian Institute of Technology-Bombay.

[9] Evans, S. Jamie and Vikram Krishnamurthy. "HMM State Estimation with Randomly Delayed Observation". IEEE Transactions on Signal Processing, Vol. 47, No. 8, Agustus, 1999.

[10] Fari, Guoliang dan Xia Xiang_Gen. "Improved Hidden Markov Models in the Wavelet-Domain". IEEE Transactions on Signal Processing, Vol. 49, No. 1, Januari, 2001.

[11] Fukada, T., K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. of ICASSP'92, vol.1, pp.137–140, 1992.

[12] Kim, Sang-Jin, Jong-Jin Kim, and Minsoo Hahn. "HMM-Based Korean Speech Synthesis System for Hand-Held Devices". IEEE Transactions on Consumer Electronics, Vol. 52, No. 4, NOVEMBER 2006.

[13] Ljolje, Andrej and E. Stephen Levinson. "Development of an Acoustic-Phonetic Hidden Markov Model for Continous Speech Recognition". IEEE Transactions on Signal Processing, Vol. 39, No. 1, Januari, 1991.

[14] Rabiner, R., and Lawrence. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". IEEE, Vol. 77, No. 2, Februari, 1989.

[15] Tokuda, K., Heiga Zen, and Alan W. Black. An HMM-Based Speech Synthesis System Applied to English. Department of Computer Science, Nagoya Institute of Technology Language Technologies Institute, Carnegie Mellon University.

[16] Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. of ICASSP 2000, vol.3, pp.1315–1318, June 2000.

[17] Tokuda, K. T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," Proc. of ICASSP, 1999.

[18] Yamagishi, Junichi. 2006. Average-Voice-Based Speech Synthesis.