TUGAS AKHIR - TF 181801

AVERAGE VOICE MODEL (AVM) BERBASIS *HIDDEN MARKOV MODEL* (HMM) PADA SINTESIS BAHASA INDONESIA

FARAHIYAH AISAH SIDIK
NRP. 02311640000055

Dosen Pembimbing:
Dr. Dhany Arifianto, S.T., M.Eng.

Departemen Teknik Fisika
Fakultas Teknologi Industri Dan Rekayasa Sistem
Institut Teknologi Sepuluh Nopember
Surabaya
2020

Halaman ini sengaja dikosongkan

FINAL PROJECT - TF 181801

AVERAGE VOICE MODEL (AVM) BASED ON HIDDEN MARKOV MODEL (HMM) IN BAHASA INDONESIA SPEECH SYNTHESIS SYSTEM

FARAHIYAH AISAH SIDIK
NRP. 02311640000055

Supervisors:
Dr. Dhany Arifianto, S.T., M.Eng

Department Of Engineering Physics
Faculty of Industrial Technology and System Engineering
Institut Teknologi Sepuluh Nopember
Surabaya
2020

Halaman ini sengaja dikosongkan

PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini.

Nama	: Farahiyah Aisah Sidik
NRP	: 02311640000055
Departemen / Prodi	: Teknik Fisika / S1 Teknik Fisika
Fakultas	: Fakultas Teknologi Industri & Rekayasa Sistem (FT-IRS)
Perguruan Tinggi	: Institut Teknologi Sepuluh Nopember

Dengan ini menyatakan bahwa Tugas Akhir dengan judul "***AVERAGE VOICE MODEL* (AVM) BERBASIS *HIDDEN MARKOV MODEL* (HMM) PADA SINTESIS BAHASA INDONESIA**" adalah benar karya saya sendiri dan bukan plagiat dari karya orang lain. Apabila di kemudian hari terbukti terdapat plagiat pada Tugas Akhir ini, maka saya bersedia menerima sanksi sesuai ketentuan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sebenarnya-benarnya.

Surabaya, 3 Agustus 2020

Yang membuat pernyataan,

Farahiyah Aisah Sidik

NRP. 02311640000055

LEMBAR PENGESAHAN

TUGAS AKHIR

AVERAGE VOICE MODEL (AVM) BERBASIS *HIDDEN MARKOV MODEL* (HMM) PADA SINTESIS BAHASA INDONESIA

Oleh:

Farahiyah Aisah Sidik

NRP. 02311640000055

Surabaya, 27 Agustus 2020

Menyetujui,

Pembimbing

Dr. Dhany Arifianto, S.T., M.Eng.

NIP. 19731007 199802 1 001

Mengetahui,

Kepala Departemen

Teknik Fisika FT-IRS ITS

Dr. Suyanto, S.T., M.T.

NIP. 19171113 199512 1 002

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

AVERAGE VOICE MODEL (AVM) BERBASIS *HIDDEN MARKOV MODEL* (HMM) PADA SINTESIS BAHASA INDONESIA

TUGAS AKHIR

Diajukan Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Teknik
pada
Program Studi S-1 Departemen Teknik Fisika
Fakultas Teknologi Industri & Rekayasa Sistem (FT-IRS)
Institut Teknologi Sepuluh Nopember

Oleh:

FARAHIYAH AISAH SIDIK

NRP. 02311640000055

Disetujui oleh Tim Penguji Tugas Akhir:

1. Dr. Dhany Arifianto, S.T., M. Eng. ..... (Pembimbing)
2. Ir. Wiratno Argo Asmoro, M.Sc. ..... (Ketua Penguji)
3. Dr. rer. nat. Ir. Aulia Muhammad, M.Sc. ..... (Penguji I)
4. Irwansyah, S.T, M.T., M. Phil., Ph.D. ..... (Penguji II)

SURABAYA

2020

Halaman ini sengaja dikosongkan

AVERAGE VOICE MODEL (AVM) BERBASIS *HIDDEN MARKOV MODEL* (HMM) PADA SINTESIS BAHASA INDONESIA

Nama	**:**	**Farahiyah Aisah Sidik**
NRP	**:**	**02311640000055**
Departemen	**:**	**Teknik Fisika FT-IRS - ITS**
Dosen Pembimbing	**:**	**Dr. Dhany Arifianto, ST, M.Eng**

ABSTRAK

Teknik *speaker adaptation* merupakan salah satu teknik pada *HMM-based speech synthesis system* (HTS) dengan kelebihan dapat mensintesis suara yang diinginkan dengan basis data yang sedikit. Pada penelitian ini dibuat sistem sintesis suara dengan teknik *speaker adaptation* berbasis HMM dengan basis data kalimat berita dan kalimat tanya. Variasi diberikan pada jumlah kalimat yang dilatih, jenis kelamin, dan jenis kalimat yang disintesis. Metode *adaptasi constrained maximum likelihood linear regression* (CMLLR) dan *maximum likelihood linear regression* (MLLR) diaplikasikan untuk dapat melakukan transformasi nilai *mean* dan *kovarian* dari parameter akustik *average voice* menjadi nilai *mean* dan *kovarian* parameter akustik dari *target speaker*. Penelitian dilakukan dengan mengguanakan 5 basis data pembicara yang dilatih kemudian terdapat 1 basis data pembicara yang akan menjadi *target speaker*. Berdasarkan pengujian objektif dengan *mel-cepstral distortion* (MCD) dan *root mean square error* (RMSE) pada log F0, diperoleh nilai MCD terbaik pembicara perempuan sebesar 11,2 pada *full training* kalimat berita dan pada pembicara laki-laki sebesar 11,1 pada *full training* kalimat berita. Nilai RMSE terbaik pembicara perempuan sebesar 0,75 pada *full training* kalimat tanya dan pembicara laki-laki sebesar 0,41 pada *full training* kalimat tanya. Berdasarkan uji subjektif, nilai MOS pada pembicara fena kalimat berita *full training* sebesar 2,75/5 dan pada pembicara mmht kalimat berita *full training* sebesar 3/5. Sehingga hasil sintesis suara dapat digolongkan “cukup baik”.

Kata Kunci: Bahasa Indonesia*, speech synthesis*, HTS*, average voice model, speaker adaptation*

Halaman ini sengaja dikosongkan

AVERAGE VOICE MODEL (AVM) BASED ON HIDDEN MARKOV MODEL (HMM) IN INDONESIA SPEECH SYNTHESIS SYSTEM

Name	**:**	**Farahiyah Aisah Sidik**
NRP	**:**	**02311640000055**
Department	**:**	**Engineering Physics FTIRS - ITS**
Supervisors	**:**	**Dr. Dhany Arifianto, ST, M.Eng**

ABSTRACT

The speaker adaptation technique is one of the techniques in HMM-based speech synthesis system (HTS) with the advantage of being able to synthesize target sound with a small database. In this research, a speech synthesis system with speaker adaptation technique based on HMM was made with a database of declarative and question sentences. Variations are given to the number of sentences being trained, the sex, and the types of sentences that are synthesized. The method of adaptation constrained maximum likelihood linear regression (CMLLR) and maximum likelihood linear regression (MLLR) was applied to be able to transform the mean and covariance of the average voice model acoustic parameters into the mean and covariance of the acoustic parameters of the target speaker. The study was conducted by using 5 database of trained speakers and then there is 1 database of speakers that will be the target speaker. Based on objective testing with mel-cepstral distortion (MCD) and root mean square error (RMSE) in log F0, obtained the best MCD value of female speakers is 11.2 in full training news sentences and in male speakers is 11.1 in full news sentence training. The best RMSE value of female speakers is 0.75 in full sentence training and male speakers is 0.41 in full sentence question training. Based on subjective tests, the MOS value for full training declarative sentences is 3/5 and for mmht speaker full training news sentences are 3/5. So the results of speech synthesis can be classified as "good enough".

Keywords **Bahasa Indonesia,** ***speech synthesis,*** **HTS,** ***average voice model, speaker adaptation***

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Puji syukur kepada Allah SWT atas limpahan rahmat, hidayah dan karunia-Nya sehingga penulis dapat menyelesaikan laporan Tugas Akhir dengan judul **"*Average Voice Model* (AVM) Berbasis *Hidden Markov Model* (HMM) Pada Sintesis Bahasa Indonesia"**. Penulis telah banyak mendapatkan bantuan dari berbagai pihak dalam menyelesaikan Tugas Akhir ini. Untuk itu penulis mengucapkan terima kasih kepada:

1. Bapak Sidik Sudarmihadi dan Ibu Fitri Fananiar selaku orangtua penulis yang selalu memberi dukungan baik secara moril dan materiil serta motivasi dan doa dalam pengerjaan Tugas Akhir ini.
2. Bapak Dr. Suyanto, S.T, M.T. selaku Ketua Departemen Teknik Fisika ITS dan Bapak Dr. Ir. Syamsul Arifin, M.T. selaku dosen wali penulis yang telah memeberikan bimbingan, serta ilmu yang sangat bermanfaat.
3. Bapak Dr. Dhany Arifianto, S.T., M.Eng., selaku dosen pembimbing yang senantiasa memberikan motivasi, bimbingan, ilmu dan arahan dalam menyelesaikan Tugas Akhir ini.
4. Bapak Dr. Suyanto, S.T, M.T. selaku Kepala Laboratorium Vibrasi dan Akustik yang telah memberikan sarana dan prasarana untuk menunjang pelaksanaan Tugas Akhir ini.
5. Teman – teman asisten Laboratorium Vibrasi dan Akustik, khususnya 2016 dan 2017 yang selalu memberikan motivasi dan dukungan
6. Teman – teman 2016, 2017,2018, dan 2019 yang tidak bisa disebutkan satu per satu.

Penulis menyadari bahwa laporan Tugas Akhir ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran dari semua pihak yang bersifat membangun selalu diharapkan demi kesempurnaan laporan ini. Harapan penulis atas laporan Tugas Akhir ini semoga bisa memberikan manfaat dan menambah wawasan bagi pembacanya.

Surabaya, 25 Agustus 2020

Penulis

Halaman ini sengaja dikosongkan

DAFTAR ISI

HALAMAN JUDUL……i
COVER PAGE..... iii
PERNYATAAN BEBAS PLAGIASI v
LEMBAR PENGESAHAN vii
LEMBAR PENGESAHAN ix
ABSTRAK xi
ABSTRACT..... xiii
KATA PENGANTAR xv
DAFTAR ISI..... xvii
DAFTAR GAMBAR xix
DAFTAR TABEL..... xxiii
BAB I PENDAHULUAN..... 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 2
1.3 Tujuan.....2
1.4 Batasan Masalah.....2
1.5 Sistematika Laporan 3
BAB II TINJAUAN PUSTAKA DAN DASAR TEORI..... 5
2.1 Bahasa Indonesia.....5
2.2 Sistem Produksi Suara.....7
2.3 *Hidden Markov Model* (HMM).....8
2.4 *HMM-based Speech Synthesis System* (HTS) 9
2.5 *Shared-Decision-Tree-based Context Clustering* 11
2.6 *Speaker Adaptation* 14

2.7 Pengujian Hasil Sintesa Suara 19
BAB III METODOLOGI PENELITIAN 21
3.1 Studi literatur 21
3.2 Instalasi dan Identifikasi HMM-*based Speech Synthesis System* (HTS) 22
3.3 HTS-*Average Voice Model* 22
3.4 HTS-*Speaker Adaptation* 23
3.5 HTS-*Speaker Adaptation* Bahasa Indonesia 26
3.6 Uji Subjektif dan Objektif 28
BAB IV HASIL DAN PEMBAHASAN 31
4.1 Hasil Segmentasi dan Labelling 31
4.2 HTS *Speaker Adaptation* Bahasa Indonesia 32
4.3 Hasil Sintesis *Average Voice Model* Bahasa Indonesia 35
4.4 Hasil Sintesis *Speaker Adaptation* Bahasa Indonesia 41
4.5 Hasil Pengujian Objektif 49
4.6 Hasil Pengujian Subjektif 55
BAB V KESIMPULAN DAN SARAN 59
5.1 Kesimpulan 59
5.2 Saran 59
DAFTAR PUSTAKA 61
LAMPIRAN 63
BIODATA PENULIS 65

DAFTAR GAMBAR

Gambar 2. 1 Sistem produksi suara manusia (Tokuda, et al., 2013) 7
Gambar 2. 2 Model sistem produksi suara dalam source-filter model (Tokuda, et al., 2013)..... 7
Gambar 2. 3 Struktur HMM (Yamagishi, 2006) 8
Gambar 2. 4 Diagram Blok HMM-based Speech Synthesis System (Yamagishi, 2006) 10
Gambar 2. 5 Vektor observasi (Tokuda, et al., 2013) 10
Gambar 2. 6 Contoh *decison tree* (Yamagishi, 2006) 11
Gambar 2. 7 Diagram blok tahap *training average voice model* (Yamagishi, 2006) 13
Gambar 2. 8 *Context clustering* pada AVM menggunakan *decision tree* untuk s*peaker dependent models* (Yamagishi, 2006)..... 13
Gambar 2. 9 *Speaker adaptation* 14
Gambar 2. 10 *Speaker adaptive training* (Yamagishi, 2006)..... 14
Gambar 2. 11 Diagram blok sintesa suara berbasis HMM menggunakan *average voice model* dan *speaker adaptation* (Yamagishi, 2006) 17
Gambar 2. 12 *Maximum likelihood linear regression* (MLLR) 18
Gambar 3. 1 Diagram alir penelitian 21
Gambar 3. 2 *Training monophone* HMMs 24
Gambar 3. 3 *Training fullcontext* HMMs..... 25
Gambar 3. 4 Pembentukan *speaker independent*..... 26
Gambar 3. 5 Proses adaptasi..... 26
Gambar 3. 6 Proses sintesis 27
Gambar 4. 1 *Full context dependent* label pada fonem 'u'..... 32
Gambar 4. 2 Grafik waktu komputasi pembicara mmht 34
Gambar 4. 3 Grafik waktu komputasi pembicara fena..... 34
Gambar 4. 4 Plot *waveform* sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b*) average voice model full training*, (c) *average voice model minimum training* 35

Gambar 4. 5 Plot frekuensi dasar sintesis suara mmht kalimat berita “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training* 37
Gambar 4. 6 Plot frekuensi dasar sintesis suara mmht kalimat tanya “benarkah ayah akan meninggalkan kami untuk waktu yang lama ?” (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training* 37
Gambar 4. 7 Plot spektral sintesis suara mmht kalimat berita “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *average voice model full training*, (c)*average voice model minimum training* 39
Gambar 4. 8 Plot spektral sintesis suara mmht kalimat tanya “benarkah ayah akan meninggalkan kami untuk waktu yang lama ?” (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training* 39
Gambar 4. 9 Plot spektral sintesis suara fena kalimat berita “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training* 40
Gambar 4. 10 Plot spektral sintesis suara fena kalimat tanya “benarkah ayah akan meninggalkan kami untuk waktu yang lama ?” (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training*, (e) *speaker adaptation minimum training* 40
Gambar 4. 11 Plot *waveform* sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training* 41
Gambar 4. 12 Plot frekuensi dasar sintesis suara mmht kalimat berita “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *speaker adaptation full training*, (c)*speaker adaptation minimum training* 43
Gambar 4. 13 Plot frekuensi dasar sintesis suara mmht kalimat tanya “benarkah ayah akan meninggalkan kami untuk waktu yang lama ?” (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*..... 44
Gambar 4. 14 Plot spektral sintesis suara mmht kalimat berita “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training* 45

Gambar 4. 15 Plot spektral sintesis suara mmht kalimat tanya “benarkah ayah akan meninggalkan kami untuk waktu yang lama ?” (a) suara asli, (b) *speaker adaptation full training,* (c) *speaker adaptation minimum training* 46
Gambar 4. 16 Plot spektral sintesis suara fena kalimat berita “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *speaker adaptation full training,* (c) *speaker adaptation minimum training* 46
Gambar 4. 17 Plot spektral sintesis suara fena kalimat tanya “benarkah ayah akan meninggalkan kami untuk waktu yang lama ?” (a) suara asli, (b) *average voice model full training,* (c) *speaker adaptation full training,* (c) *speaker adaptation minimum training* 47
Gambar 4. 18 Plot spektral perbandingan suara mmht kalimat “malam itu paman menonton televisi di kamar” (a) suara asli, (b) *speaker adaptation full training,* (c) *speaker adaptation minimum training* 48
Gambar 4. 19 Plot spektral perbandingan suara mmht kalimat “liburan kemarin aku tidak bisa pulang kampung” (a) suara asli, (b) *speaker adaptation full training,* (c) *speaker adaptation minimum training* 49
Gambar 4. 20 Hasil pengujian objektif MCD *average voice model* pada pembicara mmht 50
Gambar 4. 21 Hasil pengujian objektif MCD *average voice model* pada pembicara fena 50
Gambar 4. 22 Hasil pengujian objektif MCD *speaker adaptation* pada pembicara mmht 51
Gambar 4. 23 Hasil pengujian objektif MCD *speaker adaptation* pada pembicara fena 51
Gambar 4. 24 Hasil Pengujian objektif RMSE *average voice model* pada pembicara mmht 52
Gambar 4. 25 Hasil pengujian objektif RMSE *average voice model* pada pembicara fena 53
Gambar 4. 26 Hasil pengujian objektif RMSE speaker adaptation pada pembicara mmht 53
Gambar 4. 27 Hasil pengujian objektif RMSE *speaker adaptation* pada pembicara fena 54

Gambar 4. 28 Hasil Pengujian subjektif metode *average voice model* pada pembicara mmht 55
Gambar 4. 29 Hasil pengujian subjektif metode *average voice model* pada pembicara fena..... 56
Gambar 4. 30 Hasil pengujian subjektif metode *speaker adaptation* pada pembicara mmht 56
Gambar 4. 31 Hasil pengujian subjektif metode *speaker adaptation* pada pembicara fena..... 57

DAFTAR TABEL

Tabel 2. 1 Fonem bahasa indonesia sesuai dengan Standar *International Phonetic Alphabet* (IPA) 5
Tabel 2. 2 Kategori penilaian ACR 19
Tabel 3. 1 Tools pada sintesis suara bahasa indonesia 22
Tabel 3. 2 *Experimental set-up full training* laki-laki 27
Tabel 3. 3 *Experimental set-up full training* perempuan 27
Tabel 3. 4 *Experimental set-up minimum training* Laki-laki 28
Tabel 3. 5 *Experimental set-up minimum training* perempuan 28
Tabel 4. 1 Mono label pada kalimat "lusa aku akan pergi ke rumah paman" 31
Tabel 4. 2 Variasi *running speaker adaptation* bahasa indonesia 33
Tabel 4. 3 Waktu komputasi HTS *speaker adaptation* 34
Tabel 4. 4 Waktu komputasi HTS *average voice model* 34

Halaman ini sengaja dikosongkan

BAB I
PENDAHULUAN

1.1 Latar Belakang

Suara merupakan komponen penting dalam sistem komunikasi untuk menciptakan hubungan antar manusia dan untuk menyampaikan informasi. Kemajuan teknologi yang sangat pesat memicu adanya *human-computer communication* seperti *speech recognition*, *dialogue processing*, *speech understanding*, *natural language processing*, *speech analysis*, dan *speech synthesis*. Kemajuan teknologi tersebut dapat dimanfaatkan untuk berbagai aplikasi seperti pada sistem navigasi mobil, *e-book reader*, alat bantu suara untuk tuna laring, pengisi dialog percakapan dalam kartun, dan komunikasi robot.

Speech synthesis merupakan teknik untuk membangkitkan suara buatan dengan memasukkan teks. Untuk dapat mentransmisikan informasi yang terdapat dalam ucapan, diperlukan *speech synthesis* yang memiliki kemampuan untuk menghasilkan suara yang natural dengan prosodi dan karakter suara yang sesuai.

Pada tahun 1990-2000, *speech synthesis* berbasis *speech unit selection*, *waveform contatenation techniques*, dan *corpus concatenative speech synthesis* telah dilakukan dan menghasilkan suara yang natural dan memiliki karakter emosi yang sesuai dengan menggunakan jumlah data yang besar (Black & Campbell, 1996) (Moulines & Charpentier, 1990) (Lida, Campbell, Higuchi, & Yasumura, 2003). Ketiga teknik tersebut masih belum efektif untuk melakukan *speech synthesis* karena membutuhkan jumlah data suara yang besar, sehingga akan membutuhkan biaya yang tinggi dan waktu yang lama untuk menghasilkan suara yang natural dengan karakter emosi yang sesuai. Teknik tersebut juga akan sulit dilakukan ketika ingin menambah jumlah *speaker* untuk basis data dan *target speaker* untuk hasil output yang diinginkan.

Speech synthesis berkembang pesat dan metode *Hiddem Markov Models* (HMM) banyak digunakan karena hanya membutuhkan data suara dalam jumlah yang kecil. *HMM-based speech synthesis* (HTS) merupakan teknik *speech synthesis* yang menggunakan *Hidden Markov Model* (HMM) untuk memodelkan probabilitas

distribusi kombinasi. HTS memiliki keunggulan dalam memodifikasi model akustik suara yang dihasilkan dan hanya membutuhkan jumlah data yang kecil (Zen, Heiga, Tokuda, & Alan, 2009). *HMM-based speech synthesis system* (HTS) telah berkembang di Jepang dan Inggris. Selain itu juga telah dikembangkan teknik *speaker adaptation* dan *average voce model* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006). Metode ini secara simultan dapat mengubah karakteristik suara dan frekuensi dasar(F0) *speech synthesis* menjadi *target speaker* dengan menggunakan sejumlah basis data yang diucapkan oleh *target speaker,* dan menjadi pendekatan yang menjanjikan untuk mengatasi masalah ini. Penerapan *average voice* dan *speaker adaptatation* dalam bahasa indonesia sudah diterapkan namun masih dalam batasan karakter atau emosi yang datar (Lestari, 2018). Dalam tugas akhir ini akan dilakukan penelitian sintesis suara bahasa Indonesia berbasis *Hidden Markov Model* (HMM) dengan teknik *speaker adaptation* dan *average voice model* untuk kalimat berita dan kalimat tanya.

1.2 Rumusan Masalah

Dari paparan latar belakang diatas, maka permasalahan yang dapat diangkat pada tugas akhir ini adalah sebagai berikut :

a) Bagaimana memperoleh AVM pada sintesis alamiah bahasa indonesia berbasis *Hidden Markov Model* (HMM)?
b) Bagaimana kualitas suara yang dihasilkan oleh HTS dengan AVM?

1.3 Tujuan

Berdasarkan pemaparan latar belakang dan rumusan masalah diatas, didapatkan tujuan dari tugas akhir ini adalah sebagai berikut:

a) Memperoleh AVM sintesis bahasa Indonesia berbasis HMM
b) Mengetahui kualitas suara yang dihasilkan dari sistem *Speaker Adaptation* pada HTS menggunakan teknik AVM.

1.4 Batasan Masalah

Untuk memfokuskan penyelesaian masalah pada penelitian tugas akhir ini, diperlukan beberapa batasan masalah diantaranya sebagai berikut:

a) Suara yang digunakan untuk *training* data adalah 3 orang perempuan dan 3 orang laki-laki.
b) Jenis kalimat yang digunakan adalah kalimat tanya dan kalimat berita yang berjumlah 1529 untuk setiap speaker.
c) Hasil rekaman berupa *background noise* dan *reverberation* diabaikan
d) *Target speaker* untuk hasil sintesa suara yang diinginkan terdapat 2, yaitu 1 suara perempuan dan 1 suara laki-laki

1.5 Sistematika Laporan

Sistematika Laporan tugas akhir ini adalah sebagai berikut:

a) BAB I PENDAHULUAN

   Pada bab I ini terdiri dari latar belakang, rumusan masalah, tujuan, batasan masalah dan sistematika laporan.

b) BAB II TINJAUAN PUSTAKA DAN DASAR TEORI

   Pada bab II ini berisi tentang studi pustaka materi yang berkaitan dengan tugas akhir yang dilakukan.

c) BAB III METODOLOGI PENELITIAN

   Pada bab III ini akan dijelaskan langkah-langkah yang berkaitan dengan tugas akhir yang dilakukan.

d) BAB IV HASIL DAN PEMBAHASAN

   Pada bab IV akan dijelaskan mengenai data dari hasil sintesis suara *average voice model* dan *speaker adaptation training* berbasis HMM yang menggunakan bahasa Indonesia, analisa fitur akustik dari suara yang disintesis, dan hasil pengujian objektif serta subjektif untuk suara yang dihasilkan.

e) BAB V KESIMPULAN DAN SARAN

   Pada bab V berisi mengenai kesimpulan tentang tugas akhir yang dilakukan berdasarkan data-data yang diperoleh pada bab IV untuk menjawab rumusan masalah yang diberikan serta saran untuk pengembangan tugas akhir selanjutnya.

Halaman ini sengaja dikosongkan

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Bahasa Indonesia

Bahasa Indonesia merupakan bahasa resmi Republik Indonesia dan bahasa persatuan bangsa Indonesia. Pada dasarnya bahasa indonesia banyak menggunakan bahasa melayu tinggi (Riau). Kajian linguistik bahasa Indonesia terdiri dari beberapa tataran, yaitu tataran fonologi, morfologi, sintaks, dan leksikon. Pengucapan baku bahasa Indonesia dalam bahasa lisan diatur oleh ilmu fonologi bahasa Indonesia. Fonologi adalah ilmu yang mempelajari bunyi-bunyi bahasa yang diucapkan oleh manusia. Fonologi terdiri dari dua bagian, yaitu :

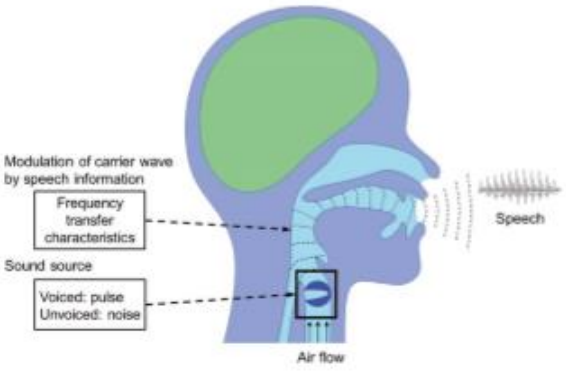
a. Fonetik adalah bagian fonologi yang mempelajari cara menghasilkan bunyi bahasa atau bagaimana suatu bunyi diucapkan oleh manusia
b. Fonemik, adalah bagian fonologi yang mempelajari ujaran menurut fungsinya sebagai pembeda arti.

Tabel 2. 1 Fonem bahasa indonesia sesuai dengan Standar *International Phonetic Alphabet* (IPA)

No.	Bahasa Indonesia	Bahasa Inggris	Contoh
1.	/a/	aa	*Father*
2.	/e/	ah, ae	*Ten*
3.	/ê/	ah, ax	*Learn*
4.	/i/	ih, iy, ix	*See, happy*
5.	/o/	ow, ao	*Got, saw*
6.	/u/	uh, uw	*Put, too*
7.	/ay/	Ay	*Five*
8.	/aw/	Aw	*Now*
9.	/ey/	Ey	*Say*
10.	/oy/	Oy	*Boy*
11.	/b/	B	*Bad*
12.	/c/	Ch	*Chain*

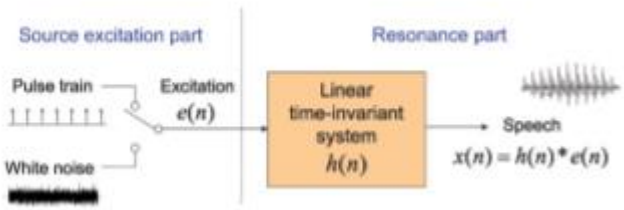
No.	Bahasa Indonesia	Bahasa Inggris	Contoh
13.	/d/	D, dx, dh	*Did*
14.	/f/	F, v	*Fall, van*
15.	/g/	G	*Got*
16.	/h/	Hh	*Hat*
17.	/j/	Jh	Jam
18.	/k/	k	*Keep*
19.	/m/	m	*Man*
20.	/l/	l	*leg*
21.	/N/	n	*no*
22.	/P/	p	*pen*
23.	/R/	r	*red*
24.	/S/	s	*so*
25.	/T/	t, th	*tea*
26.	/W/	w	*wet*
27.	/Y/	y	*yes*
28.	/Z/	z, zh	*zoo*
29.	/Kh/	-	-
30.	/Ng/	ng	*sing*
31.	/Ny/	-	-
32.	/Sy/	-	*share*
33.	sil	[]	-

2.2 Sistem Produksi Suara



Gambar 2. 1 Sistem produksi suara manusia (Tokuda, et al., 2013)

Sistem produksi suara manusia dapat dilihat pada Gambar 2.1 dimana suara dibagi menjadi tiga proses, yaitu proses pembentukan aliran udara yang berasal dari paru-paru, kemudian aliran udara melewati pita suara menjadi getaran untuk menghasilkan suara *voiced* dan *unvoiced*, dan terjadi proses artikulasi untuk menjadikan suara menjadi bunyi yang spesifik. Suara *voiced* dihasilkan saat pita suara berkontraksi, sedangkan suara *unvoiced* dihasilkan saat pita suara berelaksasi.



Gambar 2. 2 Model sistem produksi suara dalam source-filter model (Tokuda, et al., 2013)

Proses produksi yang telah digambarkan diatas merupakan implementasi dari model *source-filter*. Model *source-filter* ini menggunakan sumber berupa eksitasi dan sinyal *pulse train* untuk menghasilkan suara *voiced*, sedangkan untuk menghasilkan suara *unvoiced* digunakan *pulse train* dan *white noise*. Sinyal input tersebut kemudian difilter menggunakan filter resonansi tunggal untuk memodelkan gelombang tekanan suara akustik, yaitu *envelope* spektral dari aliran *glottal*, resonansi *vocal tract* dan efek radiasi bibir yang secara bersamaan akan menghasilkan suara. Model ini meliputi, informasi suara, frekuensi dasar (F_0), dan

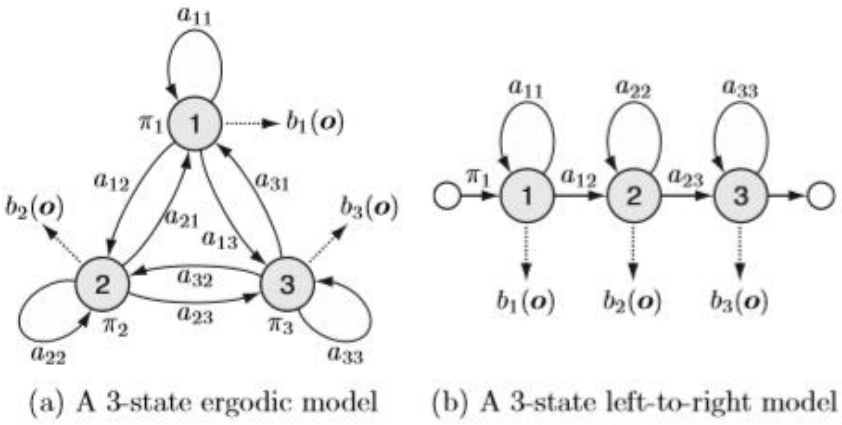
envelope spektral yang merepresentasikan koefisien *mel-cepstral* (Tokuda, et al., 2013)

2.3 *Hidden Markov Model* (HMM)

Hidden Markov Model (HMM) merupakan model statistik dalam domain waktu yang digunakan dalam sistem pengenalan suara. HMM adalah keadaan terbatas yang membangkitkan rangkain dari observasi waktu diskrit. Pada setiap unit waktu, HMM akan mengubah *state* pada proses *markov* dengan probabilitas pergeseran *state* dan membangkitkan data observasi *o* yang sesuai dengan distribusi probabilitas output pada *state* tersebut (Cahyaningtyas E. , 2015). Pada HMM, model parameter yang digunakan adalah sebagai berikut :

$$\lambda = (A, B, \Pi) \tag{2.1}$$

Struktur HMM dapat dilihat pada Gambar 2.3, struktur pada bagian kiri (a), merupakan model *3-state ergodic* yang berarti semua *state* dapat dijangkau oleh yang lain dengan sekali transisi. Struktur pada bagian kanan (b), merupakan model *left-to-right* yang menunjukkan indeks *state* akan meningkat atau tetap berdasarkan pada pertambahan waktu. Model *left-to-right* digunakan sebagai speech unit untuk memodelkan rangkaian parameter suara karena dapat memodelkan sinyal yang karakteristiknya berubah (Yamagishi, Average-Voice-Based Speech Synthesis, 2006).



Gambar 2. 3 Struktur HMM (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)

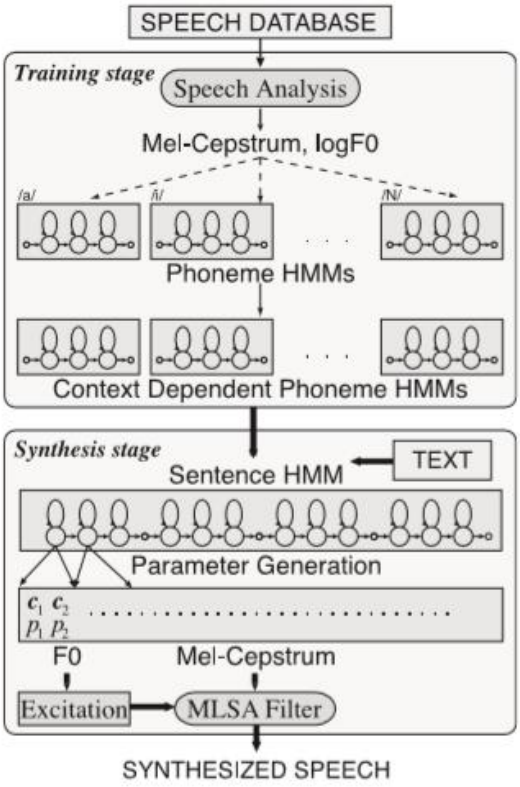
2.4 *HMM-based Speech Synthesis System* (HTS)

HMM-based speech synthesis system (HTS) merupakan teknik sintesis suara secara parametrik berbasis *Hidden Markov Model* (HMM). Keunggulan dari HTS adalah mampu untuk memodelkan dan mensintesis suara dari karakter dan emosi pembicara yang berbeda dengan menggunakan jumlah data suara yang cukup sedikit. Pada Gambar 2.4 terdapat diagram blok *HMM-based Speech Synthesis System*. Terlihat bahwa terdapat 2 proses utama dalam sistem yaitu, *training* dan *synthesis*. Pada tahapan *training*, *context dependent HMM* dilatih menggunakan basis data. *Context dependent HMM* merupakan pengenalan ucapan untuk model unit fonetik yang bergantung pada konteks unit fonetik. Parameter spektral dan F_0 diekstrasi dari setiap analisis *frame* sebagai fitur statis dari basis data dan dimodelkan menggunakan *multi-stream HMMs*, dimana distribusi keluaran bagian spektral dan F_0 dimodelkan menggunakan *continuous probability distribution* dan *multi-space probability distribution* (MSD). Pada setiap *frame* terdapat vektor observasi yang dapat dilihat pada Gambar 2.5. Kemudian, *decision tree-based context clustering* diaplikasikan secara terpisah untuk bagian spektral dan F_0 pada *context dependent phoneme* HMMs. Pada tahap pengelompokan, *decision tree* akan terbentuk berdasarkan kriteria *Minimum Description Length* (MDL).

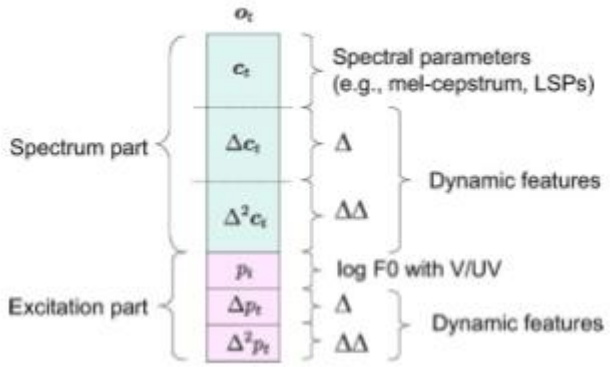
Estimasi dilakukan secara berulang untuk menghasilkan kemungkinan terbaik pada *context dependent phoneme HMMs* menggunakan algoritma *Baum-Welch*. Tahap akhir pada bagian *training* adalah penentuan durasi dari setiap fonem menggunakan distribusi *multivariate Gaussian*. Pada tahap *training*, algoritma yang digunakan adalah sebagai berikut,

$$\hat{\lambda} = \arg\max_{\lambda} P(\boldsymbol{O}|\lambda, W) \tag{2.2}$$

Pada tahap sintesis, teks yang diberikan diubah menjadi urutan label fonem yang bergantung pada *context dependent phoneme labels*. Berdasarkan urutan label, kalimat HMM dibangun dengan menggabungkan *context dependent phoneme HMMs*. Dari kalimat HMM, urutan parameter spektral dan F_0 diperoleh berdasarkan kriteria *Maximum Likelihood* (ML) dimana durasi fonem ditentukan menggunakan distribusi durasi keadaan. Kemudian, dengan menggunakan filter MLSA (Mel Log Spectral Approximation), ucapan disintesis dari sekuens urutan



Gambar 2. 4 Diagram Blok HMM-based Speech Synthesis System (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)



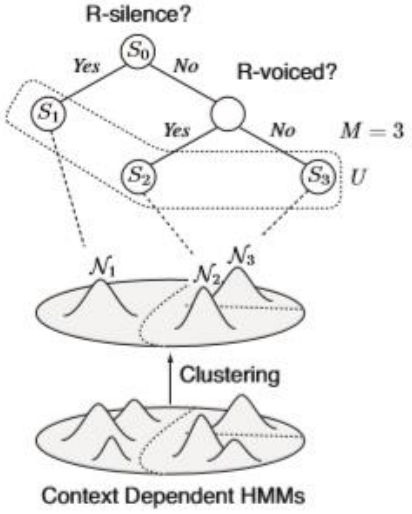
Gambar 2. 5 Vektor observasi (Tokuda, et al., 2013)

mel-cepstral dan F_0 yang dihasilkan (Yamagishi, Average-Voice-Based Speech Synthesis, 2006). Pada tahap sintesis, algoritma yang digunakan adalah sebagai berikut,

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o}|\hat{\lambda}, w) \tag{2.3}$$

2.5 *Shared-Decision-Tree-based Context Clustering*

Pada sintesis suara berbasis HMM, *speech unit* yang digunakan lebih rumit dengan mempertimbangkan konteks linguistik dan prosodi, seperti aksen frasa, pernapasan, dan informasi kalimat untuk memodelkan fitur suprasegmental dalam fitur prosodi yang tepat. Untuk mengatasi masalah ini, digunakan algoritma yang disebut *decision-tree-based-context clustering algorithm*. Algoritma ini dapat menyiapkan basis data untuk *training* dengan beragam *speech unit* dan beragam frekuensi yang bergantung pada *context dependent unit.* Pada Gambar 2,6, terdapat contoh dari sebuah *decision tree* yang merupakan *binary tree*. Masing-masing dari node memiliki pertanyaan seperti "apakah fonem sebelumnya huruf konsonan?" atau "apakah fonem selanjutnya berhenti?". Dengan menggunakan *decision-tree-based context clustering*, semua model paramater *speech unit* akan terlihat karena setiap konteks akan melewati semua *node*.



Gambar 2. 6 Contoh *decison tree* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)

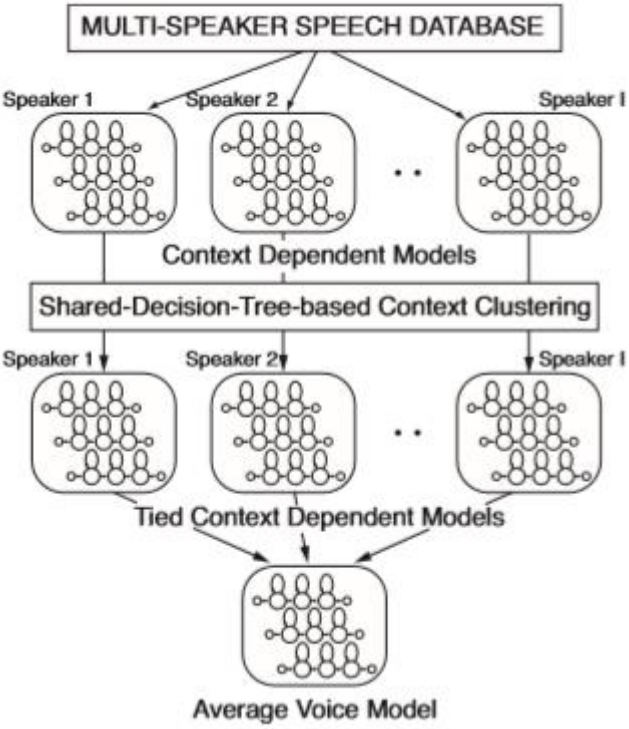
Jika dilihat pada Gambar 2.6, S_0 merupakan *node* awal dari *decision tree* dan $U(S_1,S_2,S_3,\ldots,S_M)$ merupakan model yang dihasilkan. Model set pada *node* daun dihsilkan dari *U*. PDF Gaussian N_m dihasilkan dari gabungan semua *node* S_m. *Decision-tree-based context clustering* menerapkan algoritma *Baum Welch*, dimana akan dilakukan *estimation maximization* berulang untuk mendapatkan hasil *context*

dependent HMMs terbaik, ini dapat terlihat pada tanda panah keatas pada *context dependent HMMs* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006).

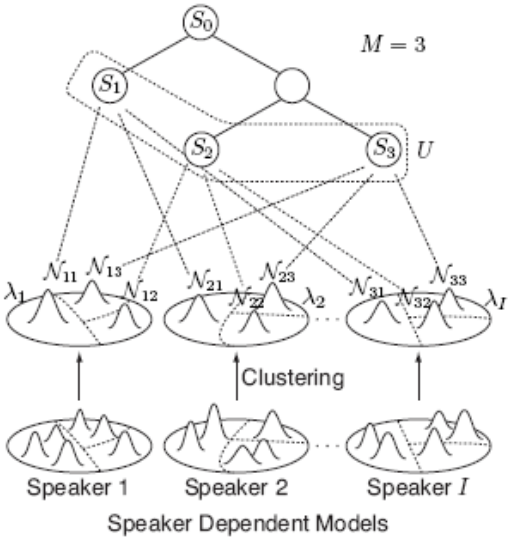
Average voice model (AVM) memiliki jumlah *training speaker* yang banyak atau lebih dari 1 untuk menghasilkan *average voice* yang akan diadaptasi untuk sintesis suara target yang diinginkan. Sistem sintesis suara menggunakan *average voice model* dapat menghasilkan sintesis suara yang mendekati dengan suara target yang diinginkan dengan menggunakan basis data suara target yang sedikit dan menggunakan metode *Maximum Likelihood Linear Regression* (MLLR). Kualitas dari suara training speaker akan mempengaruhi kualitas *average voice model* dan kualitas dari *average voice model* akan berpengaruh pada hasil sintesis suara dari model yang diadaptasi (Yamagishi, Average-Voice-Based Speech Synthesis, 2006). Basis data kalimat yang sama digunakan pada semua speaker untuk menghindari adanya perbedaan konteks. Namun, adanya banyak speaker memungkinkan adanya perbedaan karakter, konteks dan perbedaan *node* pada setiap speaker. Sehingga jika menggunakan *decision-tree-based context clustering*, tidak semua pertanyaan atau *node* akan dilalui oleh setiap speaker. Hal ini dapat menyebabkan penurunan kualitas *average voice* dan sintesa suara. Untuk mengatasi masalah ini, digunakan *shared-decision-tree-based context clustering* (STC). Dengan menggunakan metode ini, semua *node* akan dilalui oleh semua speaker (Yamagishi, Average-Voice-Based Speech Synthesis, 2006).

Pada teknik *shared-decision-tree-based context clustering* terdapat dua tahap. Tahap awal, masing-masing *speaker* melakukan *training* dengan cara sebelumnya dan membuat *decision tree* untuk *context clustering* seperti pada umumnya. Kemudian ketika decision tree terbagi, hanya pertanyaan yang berlaku untuk semua speaker yang akan diadopsi atau dipilih. Dengan menggunakan *decision tree* pada umumnya, semua *speaker* dikelompokkan dan *average voice model* akan dihasilkan dengan menggabungkan pdf Gaussian dari semua *speaker* pada setiap *node* nya.

Setelah tahap estimasi ulang pada *average voice model* dilakukan menggunakan data *training* semua *speaker*, *state duration distribution* diperoleh untuk masing-masing *speaker*. Sehingga, *state duration distribution* untuk *average voice model* dihasilkan dengan cara yang sama.

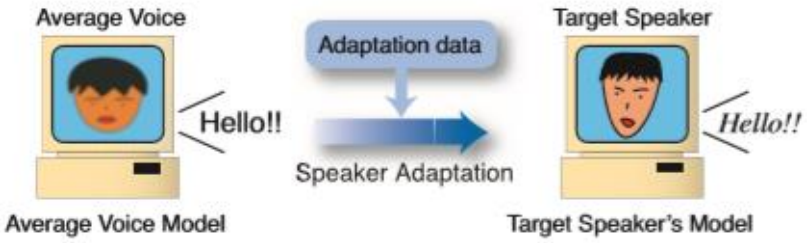


Gambar 2. 7 Diagram blok tahap *training average voice model* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)

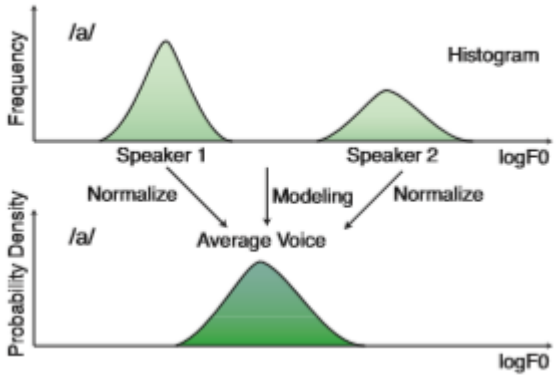


Gambar 2. 8 *Context clustering* pada AVM menggunakan *decision tree* untuk s*peaker dependent models* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)

Shared decision tree context clustering (STC) disusun berdasarkan *Minimum Description Length* (MDL). Pada Gambar 2.8, S_0 merupakan akar dari



Gambar 2. 9 *Speaker adaptation*



Gambar 2. 10 *Speaker adaptive training* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)

decision tree dan $U = (S_1, S_2, \dots, S_M)$ merupakan model dari bagian *node* kelas $\{S_1, S_2, \dots, S_M\}$. Setelah menyusun *shared decision tree*, pdfs Gaussian AVM diperoleh dengan menggabungkan pdfs Gaussian *speaker dependent models*, vektor rata-rata (μ_m) , dan matrik kovarian (Σ_m) dari pdfs Gaussian pada *node* S_m (Yamagishi, Average-Voice-Based Speech Synthesis, 2006).

$$\mu_m = \frac{\Sigma_{i=1}^{I} \Gamma_{im} \mu_{im}}{\Sigma_{i=1}^{I} \Gamma_{im}} \tag{2.4}$$

$$\Sigma_m = \frac{\Sigma_{i=1}^{I} \Gamma_m (\Sigma_{im} + \mu_{im} \mu_{im}^T)}{\Sigma_{i=1}^{I} \Gamma_{im}} \tag{2.5}$$

2.6 *Speaker Adaptation*

Speaker adaptation merupakan salah satu fitur HTS untuk mengadaptasi input suara pembicara baru (*target speaker*) tanpa harus melakukan proses *training* ulang seperti pada Gambar 2.9. Berdasarkan gambar tersebut, proses *speaker adaptation* menggunakan lebih dari satu *training speaker* sebagai input yang

kemudian dilakukan proses pelatihan dengan proses statistik menggunakan HMM untuk membentuk *average voice model* (AVM). Dengan menggunakan regresi linier, model ini dapat menghasilkan suara sesuai dengan target pembicara yang diinginkan (Yamagishi, Average-Voice-Based Speech Synthesis, 2006).

Model yang dihasilkan dari proses *training* merupakan model HMM yang memiliki fungsi probabilitas nilai *mean* dan *kovarian* antar fonemnya. Nilai **mean** dan *kovarian* merupakan representasi dari nilai parameter spektral, eksitasi, dan durasi yang telah di ekstrak dan melalui tahap *training* (Cahyaningtyas, 2018). Perbedaan antara suara *training* dengan suara rata-rata kanonik pada *speaker adaptation* diasumsikan sebagai fungsi liner vektor *mean* dari distribusi keadaan output seperti berikut,

$$\boldsymbol{\mu_i}^{(f)} = \boldsymbol{\zeta}^{(f)}\boldsymbol{\mu_i} + \boldsymbol{\epsilon}^{(f)} = \boldsymbol{Y}^{(f)}\boldsymbol{\xi}_i \tag{2.6}$$

dimana $\boldsymbol{\mu_i}^{(f)}$ dan $\boldsymbol{\mu_i}$ merupakan vektor *mean* dari distribusi keadaan output untuk *training speaker f* dan *average voice model*. $\boldsymbol{Y}^{(f)} = [\boldsymbol{\zeta}^{(f)}, \boldsymbol{\epsilon}^{(f)}]$ merupakan matriks transformasi yang menggambarkan perbedaan suara antara suara *training f* dan *average voice model* pada distribusi keadaan output. Setelah mengestimasi matriks transformasi pada masing-masing suara, dilakukan estimasi berulang pada *average voice model* sehingga transformasi pada model *training* akan maksimal .

Nilai *f* merupakan jumlah dari *training speaker*, $\boldsymbol{O} = \{\boldsymbol{O}^{(1)}, \dots, \boldsymbol{O}^{(F)}\}$ adalah semua data *training* dan $\boldsymbol{O}^{(f)} = \{\boldsymbol{O}_{1f}, \dots, \boldsymbol{O}_{Tf}\}$ merupakan data training sepanjang T_f untuk suara *f* . *Speaker adaptation* berbasis HMM secara simultan mengestimasi parameter HMM λ dan matriks transformasi $\boldsymbol{Y}^{(f)}$untuk masing-masing suara *training* untuk memaksimalkan probabilitas dari data *training* $\boldsymbol{O}$. Persamaan *speaker adaptation* berbasis HMM dengan kriteria *maximum likelihood* sebagai berikut,

$$\left(\hat{\lambda}, \tilde{\Lambda}\right) = \arg\max_{\lambda,\, \Lambda} P(\boldsymbol{O}|\hat{\lambda}, \Lambda) \tag{2.7}$$

$$= \arg\max_{\lambda,\, \Lambda} \Pi_{f=1}^{F} P(\boldsymbol{O}^{(f)}|\hat{\lambda}, \Lambda^{(f)}) \tag{2.8}$$

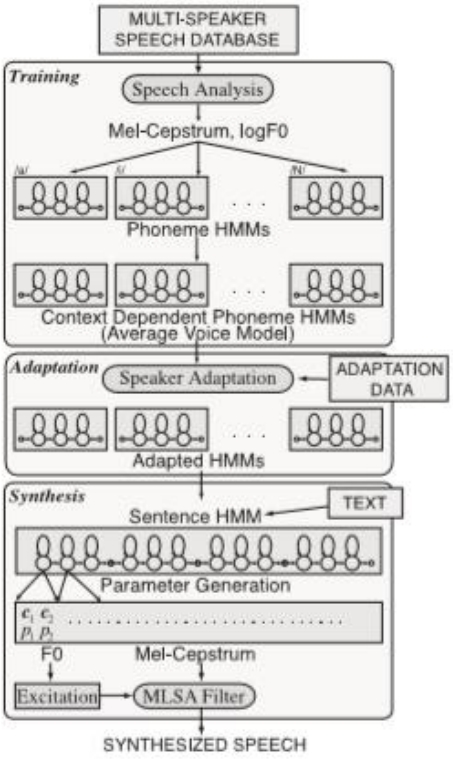
dimana $\Lambda = (W^{(1)}, \dots, W^{(F)})$ merupakan kumpulan matriks transformasi untuk *training speakers*. Terdapat tiga langkah prosedur iterasi untuk memperbarui paramater. Langkah pertama yaitu dengan mengestimasi matriks transformasi Λ dengan menetapkan nilai $\hat{\lambda}$ berdasarkan algoritma *Baum-Welch*. Vektor *mean* dari $\hat{\lambda}$ kemudian diestimasi dengan menggunakan matriks transformasi yang telah diperbarui, sedangkan matriks *kovarian* dari $\hat{\lambda}$ dijaga tetap pada nilai saat ini. Pada langkah terakhir, matriks *kovarian* dari $\hat{\lambda}$ diestimasi menggunakan matriks transformasi dan vektor *mean* yang telah diperbarui. Persamaan estimasi ulang dari kumpulan parameter $\hat{\lambda}$ untuk mendapatkan nilai vektor *mean* yang baru sebagai berikut,

$$\hat{\boldsymbol{\mu}}_i = [\boldsymbol{\Sigma}_{f=1}^{F}\boldsymbol{\Sigma}_{t=1}^{T_f}\boldsymbol{\gamma}_t(\boldsymbol{i})\tilde{\boldsymbol{\zeta}}_i^{(f)^T}\boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{\zeta}}^{(f)}]^{-1}[\boldsymbol{\Sigma}_{f=1}^{F}\boldsymbol{\Sigma}_{t=1}^{T_f}\boldsymbol{\gamma}_t(\boldsymbol{i})\tilde{\boldsymbol{\zeta}}^{(f)^T}\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{o}_{t_f} - \hat{\boldsymbol{\epsilon}}^{(f)}\right)] \quad (2.9)$$

$$\Sigma_i = \frac{\Sigma_{f=1}^{F}\Sigma_{t=1}^{T_f}\gamma_t(i)(o_{tf} - \hat{\boldsymbol{\mu}}_i^{(f)})(o_{tf} - \hat{\boldsymbol{\mu}}_i^{(f)})^T}{\Sigma_{f=1}^{F}\Sigma_{t=1}^{T_f}\gamma_t(i)} \quad (2.10)$$

dimana $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{(f)} = \zeta^{(f)}\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} + \hat{\boldsymbol{\epsilon}}^{(f)}$ merupakan vektor *mean* output yang ditransformasi menjadi *training* suara *target speaker f* menggunakan vektor *mean* output dan matriks transformasi yang telah diperbarui. $Y^{(f)} = [\hat{\zeta}^{(f)}, \hat{\epsilon}^{(f)}]$ adalah matriks transformasi yang telah diperbarui untuk *training* suara *target speaker f* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006).

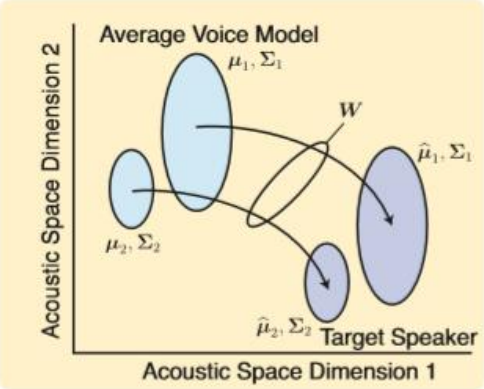
Diagram blok *speaker adaptation* terlihat pada Gambar 2.11, dimana terdapat tiga proses, yaitu proses *training*, proses adaptasi, dan proses sintesis. Proses *training* memiliki tahapan yang sama seperti pada metode sebelumnya hanya saja pada *speaker adaptation* memiliki beberapa *training speaker* dan akan menghasilkan *average voice model*. Proses adaptasi dapat mengkonversi karakteristik suara sintesis dari suara target pembicara, dilakukan adaptasi dari model awal HMMs untuk menjadi model HMMs suara target pembicara. Distribusi parameter output dari HMMs akan dimodifikasi untuk merefleksikan karakteristik suara dari target. Sehingga, parameter dan karakteristik suara sintesis yang dibangkitkan akan mendekati suara target.



Gambar 2. 11 Diagram blok sintesa suara berbasis HMM menggunakan *average voice model* dan *speaker adaptation* (Yamagishi, Average-Voice-Based Speech Synthesis, 2006)

Metode yang digunakan pada *speaker adaptation* adalah *maximum likelihood linear regression* (MLLR). MLLR merupakan transformasi linier pada parameter model untuk mendapatkan kecocokan yang lebih baik pada target speaker. MLLR melakukan transformasi dengan menghitung perubahan *mean* dan *kovarian* dari komponen campuran dengan menggunakan regresi linier. Kondisi awal dari model speaker independent atau AVM pada MLLR, seperti vektor mean dan kovarian, akan diadaptasi sesuai dengan input baru yang diberikan dengan melakukan transformasi pada mean dan kovarian parameter dengan transformasi linier (Leggetter & Woodland, 1995).

Pada Gambar 2.12 menunjukkan proses transformasi menggunakan estimasi MLLR, dari AVM menjadi *target speaker* menggunakan fungsi transformasik f_k. *Automatic Model Complexity Control* (AMCC) mengestimasi perbedaan vektor antara *target speaker* dan AVM. MLLR merupakan teknik adaptasi dengan menggunakan regresi linier pada *mean* dengan Persamaan 2.8.



Gambar 2. 12 *Maximum likelihood linear regression* (MLLR)

Constrained Maximum Likelihood Linear Regression (CMLLR) merupakan teknik adaptasi dengan menggunakan regresi linear *mean* dan *kovarian* seperti pada Persamaan 2.9 dan 2.10.

AMCC : $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \boldsymbol{\mu}_{\boldsymbol{i}} + \boldsymbol{\epsilon}_{\boldsymbol{k}}$ (2.11)

MLLR: $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \boldsymbol{\zeta}_{\boldsymbol{k}}\boldsymbol{\mu}_{\boldsymbol{i}} + \boldsymbol{\epsilon}_{\boldsymbol{k}}$ (2.12)

CMLLR: $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \boldsymbol{\zeta}_{\boldsymbol{k}}\boldsymbol{\mu}_{\boldsymbol{i}} + \boldsymbol{\epsilon}_{\boldsymbol{k}}$ (2.13)

$$\boldsymbol{\Sigma}_{\boldsymbol{i}} = \boldsymbol{\zeta}_{\boldsymbol{k}}\boldsymbol{\Sigma}_{\boldsymbol{i}}\boldsymbol{\zeta}_{\boldsymbol{K}}^{\boldsymbol{T}} \quad (2.14)$$

dimana $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$ nilai *mean* distribusi output baru dari hasil transformasi, $\boldsymbol{\mu}_{\boldsymbol{i}}$ vektor *mean* distribusi output, dan $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ nilai matriks *kovarian* dari distribusi output.

MLLR merupakan metode adaptasi suara dengan metode regresi linier. Nilai vektor *mean* diperoleh dengan melakukan transformasi linier pada nilai vektor mean hasil AVM seperti pada Gambar 2.12. Persamaan transformasi MLLR sebagai berikut,

MLLR: $b_{i_{mllr}}(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}\boldsymbol{\mu}_{\boldsymbol{i}} + \boldsymbol{\epsilon}, \boldsymbol{\Sigma}_{\boldsymbol{i}}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{Y}\boldsymbol{\xi}_{\boldsymbol{i}}, \boldsymbol{\Sigma}_{\boldsymbol{i}})$ (2.15)

CMLLR: $b_{i_{cmllr}}(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}\boldsymbol{\mu}_{\boldsymbol{i}} + \boldsymbol{\epsilon}_{\boldsymbol{i}}\boldsymbol{\zeta}_{\boldsymbol{i}}\boldsymbol{\Sigma}_{\boldsymbol{i}}\boldsymbol{\zeta}_{\boldsymbol{i}}^{\boldsymbol{T}})$

$$= \mathcal{N}(\boldsymbol{o}; \boldsymbol{Y}\boldsymbol{\xi}_{\boldsymbol{i}}\boldsymbol{\zeta}_{\boldsymbol{i}}\boldsymbol{\Sigma}_{\boldsymbol{i}}\boldsymbol{\zeta}_{\boldsymbol{i}}^{\boldsymbol{T}}) \quad (2.16)$$

dimana $\boldsymbol{\mu}_{\boldsymbol{i}}$ vektor *mean* dari distribusi output AVM, $\boldsymbol{Y} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}]$ matriks transformasi dengan dimensi *L x* (*L*+1) yang mentransformasi AVM menjadi distribusi output menjadi *target speaker*. $\boldsymbol{\xi}_{\boldsymbol{i}} = [\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{T}}, \boldsymbol{1}]^{\boldsymbol{T}}$ vektor *mean* yang diperpanjang dengan

dimensi (L+1). ζ dan ϵ adalah matriks dengan dimensi L *x* L dan vektor dengan dimensi L (Cahyaningtyas, 2018).

2.7 Pengujian Hasil Sintesa Suara

Untuk mengetahui performa dari teknik *speaker adaptation* dalam menghasilkan sintesis suara maka perlu dilakukan evaluasi untuk mengetahui tingkat kenaturalan suara yang dihasilkan. Salah satu teknik untuk mengevaluasi performa dari sistem sintesis suara adalah dengan metode subjektif dan objektif.

2.7.1 Metode Pengujian Subjektif

Metode subjektif yang digunakan untuk mengevaluasi tingkat kenaturalan hasil sintesa suara adalah metode *Mean Opinion Score* (MOS). Metode MOS digunakan untuk mengukur kualitas dari suatu sistem dengan menggunakan nilai subjektif yang ada dalam rekomendasi ITU-T P.800. Pengujian dilakukan dengan memperdengarkan suara yang akan dinilai, kemudian naracoba akan memberikan nilai dengan *Absolute Category Rating* (ACR). ACR tes menggunakan lima kategori penilaian dari nilai 1-5 seperti pada Tabel 2.2. Kemudian hasil dari nilai ACR akan di rata-rata untuk mendapatkan nilai MOS.

Tabel 2. 2 Kategori penilaian ACR

Kategori	Nilai
Sangat bagus	5
Bagus	4
Sedang	3
Buruk	2
Sangat buruk	1

2.7.2 Metode Pengujian Objektif

Metode pengujian objektif digunakan untuk menilai kualitas suara dari beberapa parameter sinyalnya, misalnya dilihat dari bentuk sinyal dalam domain waktu atupun dalam domain frekuensinya. Salah satu metode objektif yang digunakan untuk menilai kualitas suara adalah dengan menggunakan *Mel-Cepstral Distortion* (MCD) dan *Root Mean Square Error* (RMSE). MCD merupakan suatu ukuran yang menunjukkan distorsi dari *Mel-Frequency Cepstral Coefficient* (MFCC). Ukuran inilah yang digunakan sebagai dasar untuk menentukan nilai

kualitas suara yang dihasilkan dari sebuah sistem sintesis suara. Nilai MCD dapat diperoleh melalui persamaan berikut ini,

$$MCD = 10/ln10 \sqrt{2 \sum_{i=1}^{24} \left(mc_i^{(t)} - mc_i^{(c)} \right)^2} \quad (2.19)$$

dengan $mc_i^{(t)}$ merupakan nilai MFCC yang digunakan sebagai acuan dan $mc_i^{(c)}$merupakan nilai MFCC yang diprediksi.

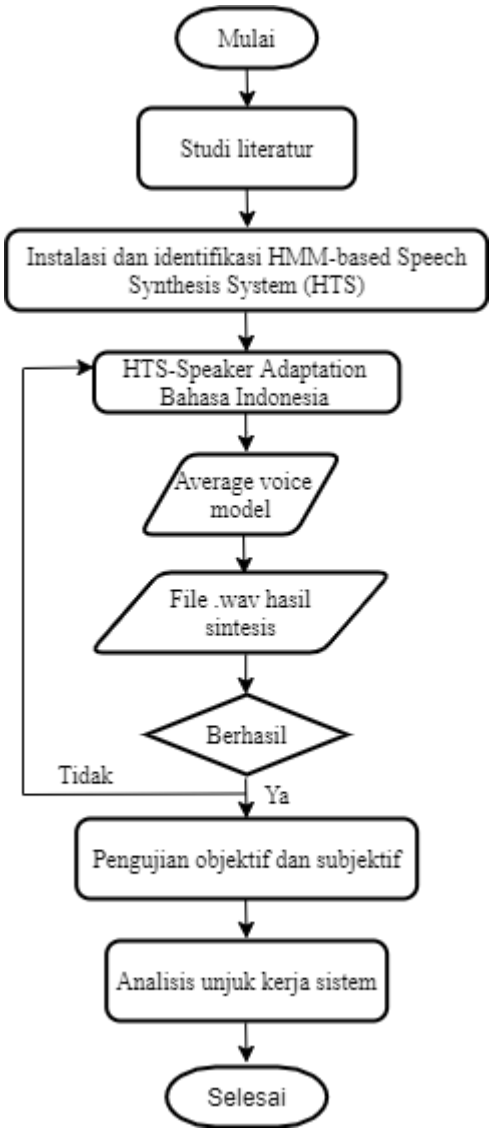
Metode R*oot Mean Square Error* (RMSE) mengukur berapa besar nilai *error* dari dua buah *dataset*. RMSE digunakan untuk membandingkan nilai sinyal yang ditinjau dengan sinyal model (*baseline*) dalam domain waktu. Nilai RMSE dapat diperoleh dari persamaan (2.20).

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(X_{obs,i} - X_{model,i} \right)^2} \quad (2.20)$$

dimana $X_{obs,i}$ merupakan sinyal yang ditinjau dan $X_{model,i}$ merupakan sinyal model pada waktu ke *i*. Nilai RMSE yang semakin kecil menunjukkan sinyal yang ditinjau semakin mirip dengan sinyal modelnya.

BAB III
METODOLOGI PENELITIAN

Pada bab ini akan dijelaskan metode yang digunakan untuk menyelesaikan permasalahan yang telah dituliskan pada BAB I. Diagram alir yang digunakan pada penelitian ini dapat dilihat pada Gambar 3.1



Gambar 3. 1 Diagram alir penelitian

3.1 Studi literatur

Berdasarkan hasil studi literatur, didapatkan teori penunjang terhadap metode yang akan dilakukan yang dapat dilihat pada Bab II. Jumlah fonem pada bahasa

Indonesia sebanyak 33. Kemudian, akan diterapkan metode statistik *Hidden Markov Model* (HMM) untuk pengenalan suara. Sintesis suara berbasis HMM dengan teknik *speaker adaptation* memiliki tiga tahap, yaitu tahap *training*, adaptasi dan tahap sintesis. Pada metode *Average Voice Model* (AVM) akan digunakan *shared-decision-tress-based context clustering* untuk mendapatkan suara *average voice* yang baik. Pada tahap adaptasi speaker, hanya diperlukan basis data training yang sedikit dan digunakan *Maximum Likelihood Linear Regression* (MLLR) dan *Constrained Maximum Likelihood Linear Regression* (CMLLR) untuk dapat menyesuakan nilai *mean* dan *kovarian* dari suara *average voice* ke suara target yang diinginkan.

3.2 Instalasi dan Identifikasi HMM-*based Speech Synthesis System* (HTS)

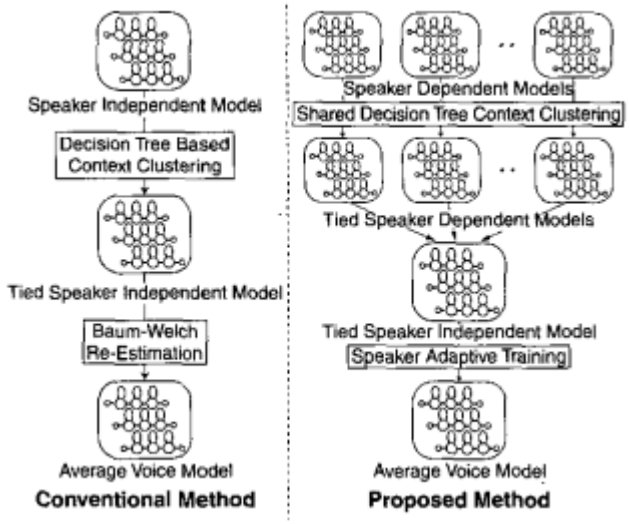
HMM-based Speech Synthesis System (HTS) dipublikasikan pertama kali sebagai *open source software* yang merupakan perluasan dari *HMM toolkit* (HTK) pada tahun 2002 oleh kelompok kerja HTS. HTS memiliki banyak program dan *tools* untuk melakukan proses *training*, adaptasi dan sintesis. Program dan *tools* yang digunakan pada HTS untuk melakukan sintesis suara dapat dilihat pada Tabel 3.1.

Tabel 3. 1 Tools pada sintesis suara bahasa indonesia

Keterangan	***Tools***
HMM training, adaptasi dan sintesis proses	HMM *Toolkit* (HTK)-3.4.1
	HMM-*based speech synthesis system* (HTS)-2.3
	HTS_engine_API-1.10
Signal processing	*Speech Signal Processing Toolkit* (SPTK) -3.10
	Edinburgh Speech Tools
Front End	*Festival Speech Synthesis System*

3.3 HTS-*Average Voice Model*

Average voice model memiliki struktur dasar yang sama dengan dengan HTS pada umumnya, hanya saja *average voice model* merupakan salah satu tahap pada teknik *speaker adaptation* yang dihasilkan pada proses training sebelum proses adaptasi seperti pada Gambar 3.2.



Gambar 3. 2 Diagram blok *training average voice model*

Pada proses *training*, basis data dari *multi-speaker* akan dilatih masing-masing untuk mendapatkan *speaker dependent phoneme HMMs*. Model akustik yang didapatkan dari masing-masing parameter akustik speaker seperti spektrum dan F_0 dimodelkan menggunakan *continuous probability distribution*. Untuk mendapatkan hasil yang tidak bias pada semua *speaker* yang dilatih, maka masing-masing-masing speaker dilatih kembali menggunakan *shared decision tree context clustering* untuk mendapatkan *speaker independent phoneme HMMs*. Parameter-parameter akustik antar speaker yang dihasilkan akan digabungkan menggunakan *multi-space distribution* dan *speaker adaptive training*. Sehingga, didapatkan model akustik gabungan dari semua *speaker* dengan nilai bias yang kecil yang dapat disebut model parameter akustik *average voice model* (Yamagishi, Masuko, Tokuda, & Kobayashi, 2003).

3.4 HTS-*Speaker Adaptation*

Speaker adaptation memiliki tiga tahapan, yaitu *training*, adaptasi, dan sintesis. Namun sebelum memulai proses *training*, dilakukan persiapan basis data. Persiapan basis data merupakan input untuk menjalankan *training*. Jumlah persiapan basis data disesuaikan dengan jumlah pembicara yang digunakan.

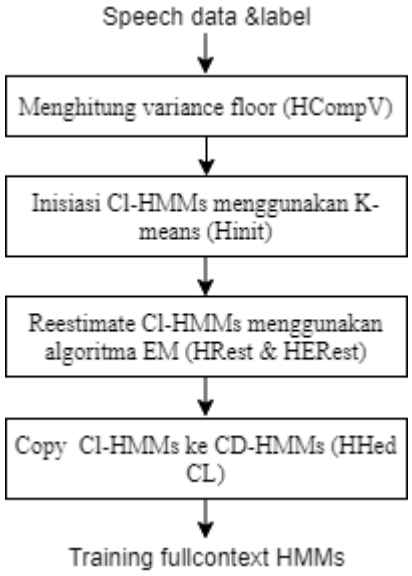
a. Persiapan Data

Adapun persiapan basis data diantaranya sebagai berikut:

- File .raw, merupakan format audio yang dihasilkan dari konversi suara rekaman .wav. Konversi ini bertujuan untuk menghilangkan informasi *header* pada format file sehingga akan lebih mudah untuk diolah
- File .utt, merupakan file informasi teks dari kalimat yang diucapkan oleh setiap rekaman.
- File .lab, merupakan informasi label kalimat untuk proses *training*, adaptasi dan sintesis. Label dilakukan pada setiap fonem pada suatu kalimat berdasarkan waktu pengucapan fonem tersebut. Sehingga pada file .lab akan diketahui durasi dari pengucapan setiap fonemnya, dan informasi prosodi pada setiap kalimat yang diucapkan.
- File *question*, merupakan file informasi pohon keputusan untuk membuat pemodelan semua parameter akustik dari basis data bahasa indonesia yang berdasarkan kaidah fonem yang berlaku.
- File .txt, merupakan kalimat *prompt* dari semua basis data.

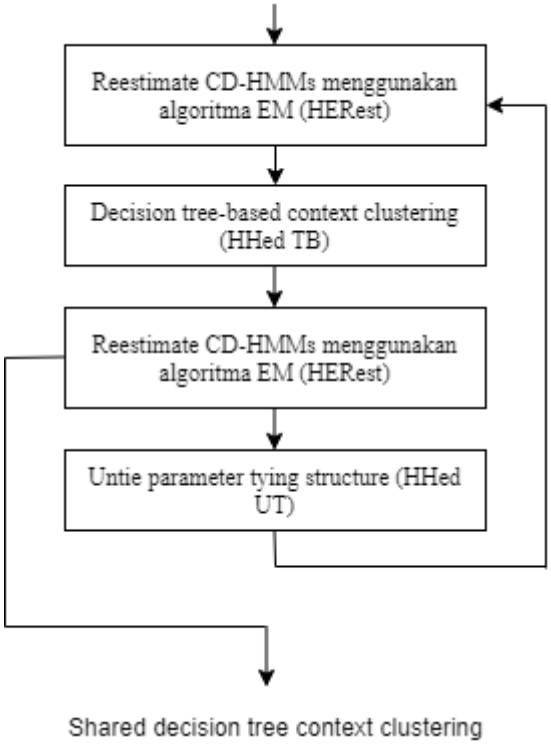
b. Proses *Training*

Proses *training* dilakukan untuk mendapatkan model paremeter akustik suara dari suara basis data. Model parameter akustik ini digunakan untuk menghasilkan fitur suara. Lamanya proses *training* bergantung pada jumlah pembicara yang digunakan dan jumlah basis data yang digunakan.



Gambar 3. 3 *Training monophone* HMMs

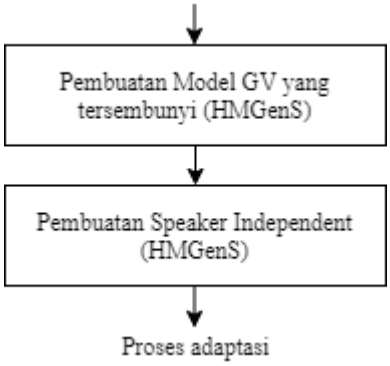
Pada Gambar 3.2, dapat dilihat proses *training monophone* merupakan tahap awal. *Monophone* dibentuk berdasarkan masing-masing fonem yang kemudian membentuk parameter akustik. Selanjutnya *monophone* diestimasi berulang dengan menggunakan *tools* "*HRest* dan *HERest*". Kemudian digunakan *tools* "*HHed CL*" untuk meningkatkan keakurataan dari *monophone*.



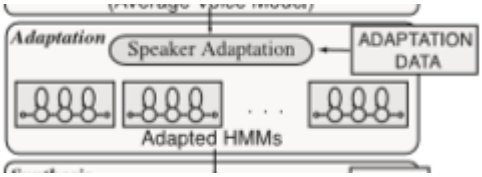
Gambar 3. 4 *Training fullcontext* HMMs

Pada Gambar 3.3 menunjukkan langkah dan *tools* yang digunakan dalam pembentukan *fullcontext* HMM. Parameter akustik akan diestimasi ulang dengan menggunakan *tools* "*HERest*". Kemudian pemodelan parameter akustik akan dibentuk menggunakan pohon keputusan dengan *tools* "*HHed TB*". Parameter akustik yang memiliki kesamaan kemudian di estimasi ulang dengan *tools* "*HERest*". Hasil dari estimasi tersebut akan diuraikan dengan menggunakan *tools* "*HHed UT*".

Pada Gambar 3.4 terdapat proses pembentukan *speaker independent.* Dimana pada proses ini akan ada tahap *shared decision tree* yang kemudian akan menghasilkan *speaker independent* atau *average voice*.



Gambar 3. 5 Pembentukan *speaker independent*



Gambar 3. 6 Proses adaptasi

c. Proses Adaptasi

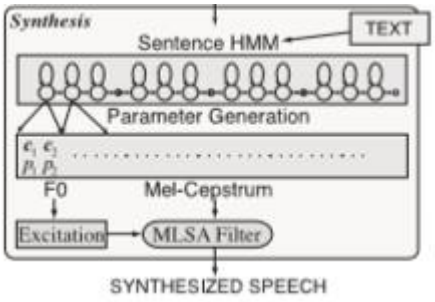
Pada proses adaptasi, akan ada input basis data pembicara baru yang akan menjadi objek untuk diadaptasi. Adaptasi dilakukan menggunakan metode MLLR dengan mentransformasikan nilai *mean* dan *kovarian average voice* menjadi suara yang diadaptasi. Pada proses adaptasi, menggunakan *tools* "*HHEd* dan *HERest*" pada HTS.

d. Proses Sintesis

Pada proses sintesis, input text dalam bentuk.lab akan dibentuk model parameter akustiknya dengan metode *maximum likelihood probability*. Kemudian untuk menghasilkan sintesis suara sesuai dengan input yang diberikan, digunakan filter *mel log spectrum approximation* (MLSA). Sintesis suara yang dihasilkan merupakan suara *average voice* dan suara *speaker adaptation*.

3.5 HTS-*Speaker Adaptation* Bahasa Indonesia

Pada penelitian ini, digunakan teknik *speaker adaptation* dalam sintesis suara. Basis data merupakan bahasa indonesia dimana terdapat 6 pembicara. Kemudian, untuk kalimat yang disintesi merupakan kalimat tanya dan kalimat



Gambar 3. 7 Proses sintesis

berita. Terdapat 4 eksperimen *set up* pada penelitian ini seperti pada Tabel 3.2 - Tabel 3.5

Tabel 3. 2 *Experimental set-up full training* laki-laki

Demo	HTS_demo_cmu_arctic_adapt
Database	Vibid
Jenis kalimat	Kalimat berita dan tanya
Training speaker	Laki-laki : mjra, meia Perempuan : fala, fbap, fena
Adaptation speaker	Mmht
Training data	1529
Adaptation data	1529
Test data	100
Sampling rate	16 KHz
Frame length	25 ms
Frame shift	5 ms
F0_ranges	fala: 125-610 Hz, meia: 80-260 Hz fbap: 185-540 Hz, mjra: 75-295 Hz fena: 145-325 Hz , mmht: 70-160 Hz

Tabel 3. 3 *Experimental set-up full training* perempuan

Demo	HTS_demo_cmu_arctic_adapt
Database	Vibid
Jenis kalimat	Kalimat berita dan tanya
Training speaker	Laki-laki : mjra, meia, mmht Perempuan : fala, fbap
Adaptation speaker	Fena
Training data	1529
Adaptation data	1529
Test data	100
Sampling rate	16 KHz
Frame length	25 ms
Frame shift	5 ms
F0_ranges	fala: 125-610 Hz, meia: 80-260 Hz

	fbap: 185-540 Hz, mjra: 75-295 Hz fena: 145-325 Hz , mmht: 70-160 Hz

Tabel 3. 4 *Experimental set-up minimum training* Laki-laki

Demo	HTS_demo_cmu_arctic_adapt
Database	Vibid
Jenis kalimat	Kalimat berita dan tanya
Training speaker	Laki-laki : mjra, meia Perempuan : fala, fbap, fena
Adaptation speaker	Mmht
Training data	120
Adaptation data	120
Test data	100
Sampling rate	16 KHz
Frame length	25 ms
Frame shift	5 ms
F0_ranges	fala: 125-610 Hz, meia: 80-260 Hz fbap: 185-540 Hz, mjra: 75-295 Hz fena: 145-325 Hz , mmht: 70-160 Hz

Tabel 3. 5 *Experimental set-up minimum training* perempuan

Demo	HTS_demo_cmu_arctic_adapt
Database	Vibid
Jenis kalimat	Kalimat berita dan tanya
Training speaker	Laki-laki : mjra, meia, mmht Perempuan : fala, fbap
Adaptation speaker	Fena
Training data	120
Adaptation data	120
Test data	100
Sampling rate	16 KHz
Frame length	25 ms
Frame shift	5 ms
F0_ranges	fala: 125-610 Hz, meia: 80-260 Hz fbap: 185-540 Hz, mjra: 75-295 Hz fena: 145-325 Hz , mmht: 70-160 Hz

3.6 Uji Subjektif dan Objektif

Uji subjektif yang dilakukan menggunakan MOS seperti yang sudah dijelaskan pada bagian teori penunjang. Penelitian ini melakukan uji subjektif dengan 16 orang naracoba. Hasil suara yang diujikan berupa suara asli dan suara hasil sintesis. Jumlah kalimat yang diperdengarkan kepada naracoba adalah 10

kalimat referensi dan 10 kalimat uji untuk masing-masing eksperimen. Waktu yang diperlukan naracoba untuk mendengarkan kalimat sekitar 60 menit.

Uji objektif mengacu pada penelitian yaitu mengukur kualitas *speech*. Metode uji yang digunakan dalam penelitian ini adalah *Mel-cepstral Distortion* (MCD) untuk mengetahui besar distorsi dari *Mel-Frequency Cepstral Coefficient* (MFCC). Metode uji lainnya yaitu *Root Mean Square Error* (RMSE) yang digunakan untuk mengukur variasi dari F_0 antara suara asli dan suara hasil sintesis.

BAB IV
HASIL DAN PEMBAHASAN

Berdasarkan beberapa tahapan yang telah berhasil dilakukan sesuai dengan prosedur pada BAB III, didapatkan hasil sebagai berikut :

4.1 Hasil Segmentasi dan Labelling

Pada tahap awal pemrosesan data, akan dilakukan segmentasi dan labelling secara otomatis menggunakan *tools festival* dan *festvox*. Segmentasi dan labelling dilakukan pada 1529 kalimat pada 6 pembicara. Segmentasi suara rekaman dilakukan berdasarkan kaidah fonem bahasa indonesia dan label berdasarkan waktu dari pengucapan masing-masing fonem. Penelitian ini dalam melakukan labeling menggunakan metode *context-dependent-label* yang secara otomatis dijalankan saat proses persiapan oleh HTS. Terdapat dua jenis *context-dependent-label* yang digunakan yaitu *mono-labeling* dan *full-context labeling*.

Pada *mono-labeling*, hanya terdapat informasi fonem beserta durasi fonem dalam satu kalimat. Pada Tabel 4.1 menunjukkan label mono dengan kalimat “lusa aku akan pergi ke rumah paman” oleh pembicara fala. Pada kalimat tersebut memiliki 30 fonem dan durasi total 34100001 ns atau 0,03 s. Kemudian, label mono ini akan digunakan untuk membuat *full-context labelling.*

Tabel 4. 1 Mono label pada kalimat "lusa aku akan pergi ke rumah paman"

Mulai (ns)	Akhir (ns)	Fonem
0	1100000	SIL
1100000	2200000	l
2200000	3300000	u
3300000	4400000	s
4400000	5500000	a
5500000	6600000	a
6600000	7700000	k
7700000	8800000	u
8800000	9900000	a
9900000	11000000	k
11000000	12100000	a
12100000	13200001	n
13200001	14299999	p
14299999	15400000	e
15400000	16500000	r

Mulai (ns)	Akhir (ns)	Fonem
16500000	17600000	g
17600000	18700000	i
18700000	19800000	k
19800000	20899999	e
20899999	22000000	r
22000000	23099999	u
23099999	24200001	m
24200001	25300000	a
25300000	26400001	h
26400001	27500000	p
27500000	28599999	a
28599999	29700000	m
29700000	31350000	a
31350000	33000000	n
33000000	34100001	SIL

2200000 3300000	SIL^l-u+s=a	@2_1/A:x_x_x/B:1-1-2@1-2&1-12#0-1$0-1!x-8…
Durasi Fonem	Fonem	Informasi prosody

Gambar 4. 1 *Full context dependent* label pada fonem 'u'

Full context label menggunakan format HTS yang memiliki 53 macam konteks untuk mengekspresikan durasi fonem, fonem dan informasi prosody yang berkaitan dengan intonasi dari pengucapan kalimat seperti pada Gambar 4.1. Pada Gambar 4.1 menunjukkan *full context label* untuk fonem “u”, dimana pada label tersebut terdapat informasi durasi fonem dan informasi prosody dari fonem “u” pada kalimat “lusa aku akan pergi ke rumah paman”.

Informasi yang ada pada *context-dependent-label* meliputi :

- *{sebelum, selanjutnya} dua fonem*
- *Posisi fonem saat ini pada suku kata saat ini*
- *Fonem pada suku kata {sebelum, saat ini, selanjutnya}*
- *{penekanan} pada suku kata {sebelum, saat ini, selanjutnya}*
- *Posisi dari suku kata saat ini pada kata saat ini*
- *# pada {penekanan} suku kata {sebelum, selanjutnya} pada satu frasa*
- *# pada {penekanan} suku kata dari suku kata {sebelum, selanjutnya}*
- *…*

4.2 HTS *Speaker Adaptation* Bahasa Indonesia

Speaker adaptation merupakan salah satu fitur HTS untuk mengadaptasi input suara pembicara baru (*target speaker*) tanpa harus melakukan proses *training* ulang, *Speaker adaptation* memiliki 3 proses yaitu proses *training*, adaptasi dan sintesis. Pada *speaker adaptation* bahasa indonesia ini telah dilakukan *training* pada 3 pembicara perempuan (fala, fbap, fena) dan 3 pembicara laki-laki (mmht, meia, mjra) dengan masing-masing kalimat 1529 yang termasuk kalimat tanya dan kalimat berita. Dari rekaman suara 6 pembicara tersebut akan dibentuk basis data ujaran (*speech corpus*) yang meliputi file suara “,raw”, file label “.lab”, dan file transkripsi “.utt” serta penyesuaian *question* bahasa indonesia. Pada tahap *training* digunakan metode minimum description length (MDL) untuk menghilangkan pengaruh karakter pembicara yang berbeda dari seluruh pembicara training

(Yamagishi, Average-Voice-Based Speech Synthesis, 2006). Pada proses training, dilakukan esperimental *set-up* seperti pada Tabel 3.2-3.5. Kemudian, akan dihasilkan model akustik rata-rata dari 6 pembicara, sehingga akan terbentuk *average voice model*.

Pada tahap adaptasi, akan diinput 1 *target speaker* baru dengan 1529 kalimat yang sama dengan pembicara pada tahap *training*. Basis data *target speaker* akan melalui segmentasi, labelling, dan *clustering* hingga mendapatkan model paramater akustik untuk *target speaker*. Kemudian akan diaplikasikan metode *maximum likelihood linear regression* (MLLR) dan *maximum a posterior*i (MAP) pada *average voice model* (AVM). Metode MLLR dan MAP dapat melakukan tranformasi menggunakan regresi linier pada nilai *mean* dan *kovarian* dari parameter akustik *average voice model* (AVM) pada kondisi awal, kemudian akan diaplikasikan metode *constrained maximum likelihood linear regression* (CMLLR) untuk melanjutkan transformasi parameter akustik menggunakan regresi linier pada matriks *mean* dan *kovarian* menjadi sesuai dengan nilai *mean* dan *kovarian* fitur akustik *target speaker*. Kemudian, CMLLR akan melakukan distribusi durasi untuk target speaker pada saat yang bersamaan.

Penelitian ini akan dilakukan dengan menggunakan dua variasi jumlah kalimat untuk *training speaker* dan *target speaker* yaitu 1529 dan 120 dengan suara yang akan diadaptasi adalah suara perempuan (fena) dan suara laki-laki (mmht) dengan dua jenis intonasi yaitu kalimat berita dan kalimat tanya seperti pada Tabel 4.2.

Tabel 4. 2 Variasi *running speaker adaptation* bahasa indonesia

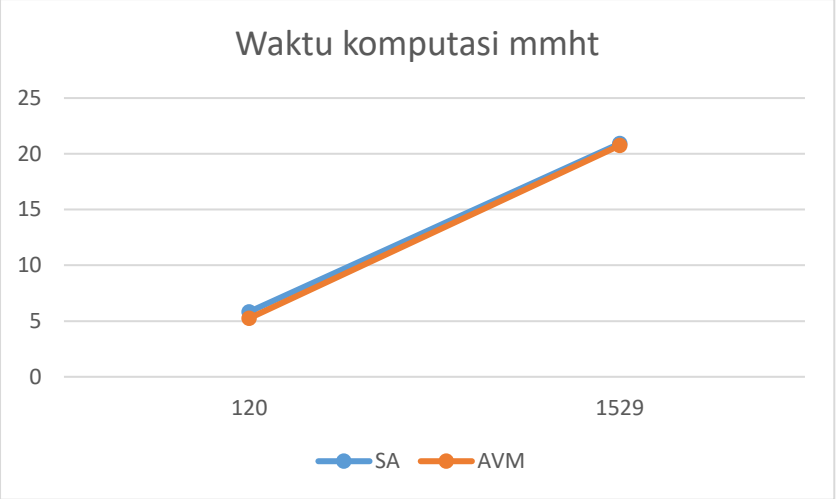
Jumlah Kalimat *Training*	Jumlah Kalimat Adaptasi	*Training Speaker*	*Target Speaker*	Kalimat Adaptasi
1529	1529	fala, fbap, fena, meia, mjra	mmht	kalimat tanya & kalimat berita
120	120	fala ,fbap, fena, meia, mjra	mmht	kalimat tanya & kalimat berita
1529	1529	fala, fbap, meia, mjra, mmht	fena	kalimat tanya & kalimat berita
120	120	fala, fbap, meia, mjra, mmht	fena	kalimat tanya & kalimat berita

Tabel 4. 3 Waktu komputasi HTS *speaker adaptation*

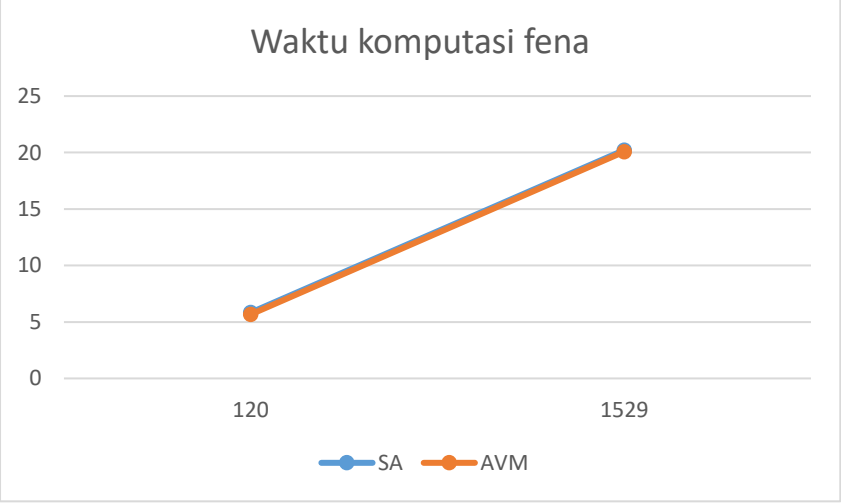
Jumlah Kalimat Training	Waktu (jam)			
	Tanya		Berita	
	mmht	fena	mmht	fena
120	05:24:26	05:49:27	05:28:44	05:51:53
1529	20:58::55	20:10:32	20:50:03	20:15:44

Tabel 4. 4 Waktu komputasi HTS *average voice model*

Jumlah Kalimat Training	Waktu (jam)			
	Tanya		Berita	
	mmht	fena	mmht	fena
120	05:19:22	05:37:25	05:13:01	05:41:49
1529	20:46:47	20:01:17	20:50:31	20:07:13



Gambar 4. 2 Grafik waktu komputasi pembicara mmht

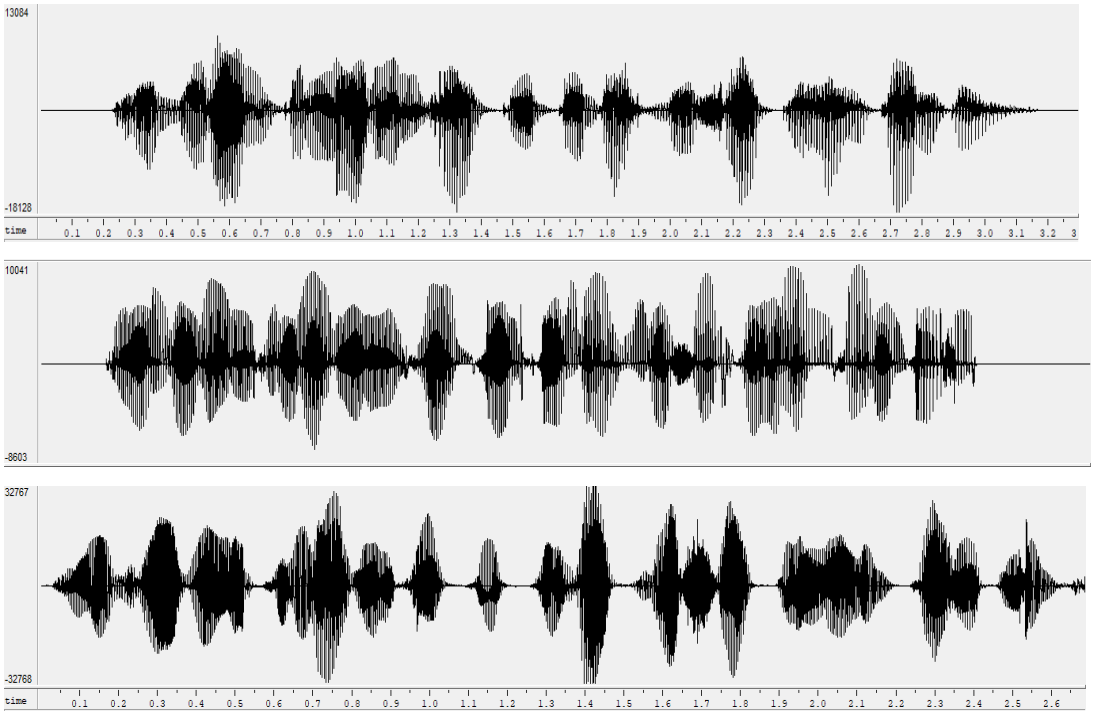


Gambar 4. 3 Grafik waktu komputasi pembicara fena

Pada Tabel 4.3-4.4 serta Gambar 4.2-4.3 dapat dilihat lama waktu yang dibutuhkan untuk menjalankan program sintesis suara dengan variasi pada jumlah kalimat trainingnya. Semakin bertambah jumlah kalimat training maka akan semakin banyak waktu yang dibutuhkan untuk melakukan komputasi.

4.3 Hasil Sintesis *Average Voice Model* Bahasa Indonesia

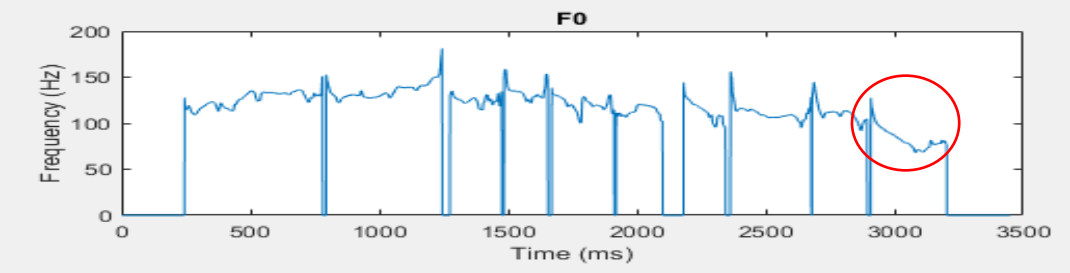
Model akustik yang dihasilkan pada tahap *training* merupakan model akustik *average voice model* dari lima *speaker* pada basis data *training*. Hasil model akustik *average voice model* akan disintesis pada proses sintesis. Pada tahap sintesis, kalimat yang disintesis melewati segmentasi, labeling, dan proses HMM model akustik untuk membentuk paramater akustik untuk kalimat-kalimat tersebut. Sehingga, ketika melewati MLSA filter akan menjadi suara dengan format file ".wav". Identifikasi suara sintesis dapat dilihat pada plot *waveform*, frekuensi dasar, dan spektral yang menunjukkan intonasi dari kalimat, kejelasan pada setiap kata, dan susunan antar fonem pada kalimat. *Waveform* untuk sintesis suara *average voice model* mmht kalimat berita terdapat pada Gambar 4.2, Plot *waveform* dihasilkan menggunakan *software* wavesurfer.

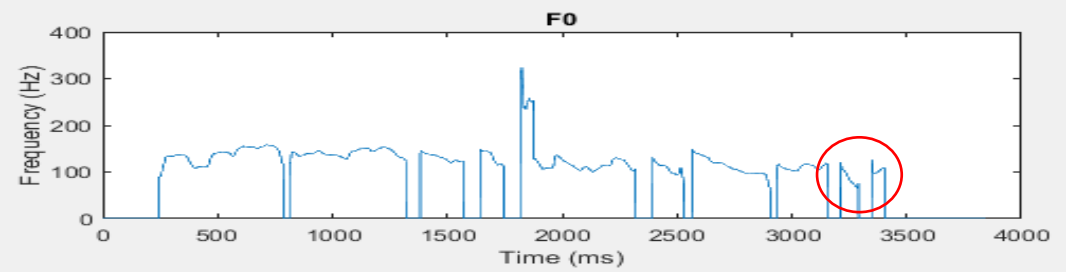


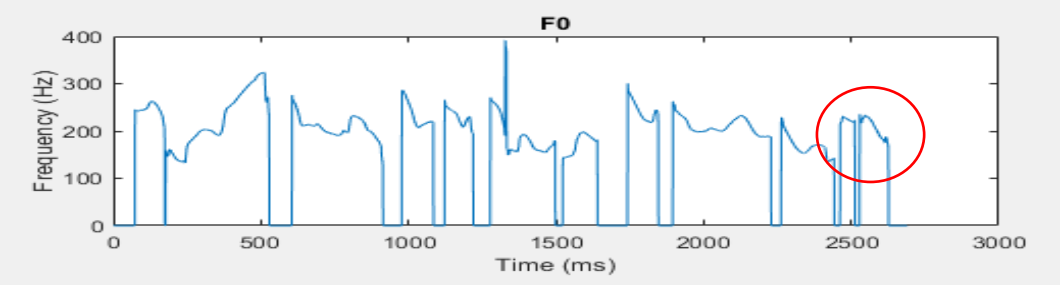
Gambar 4. 4 Plot *waveform* sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b*) average voice model full training*, (c) *average voice model minimum training*

Pada Gambar 4.4 merupakan plot perbandingan *waveform* suara asli dan suara hasil sintesis *average voice model* kalimat berita pada pembicara mmht. Sumbu-x menunjukkan informasi waktu dan sumbu-y menunjukkan informasi amplitudo sinyal suara. Pada gambar tersebut terlihat jelas terdapat perbesaran amplitudo pada *waveform* hasil sintesis *average voice model* yang disebabkan karena hasil amplitudo rata-rata lima *speaker*. Kemudian terdapat perbedaan durasi yang disebabkan oleh bedanya durasi pengucapan dari masing-masing *speaker*. Sehingga saat suara hasil sintesis didengar, terdengar jelas kalimatnya namun tidak memiliki karakteristik suara yang jelas.

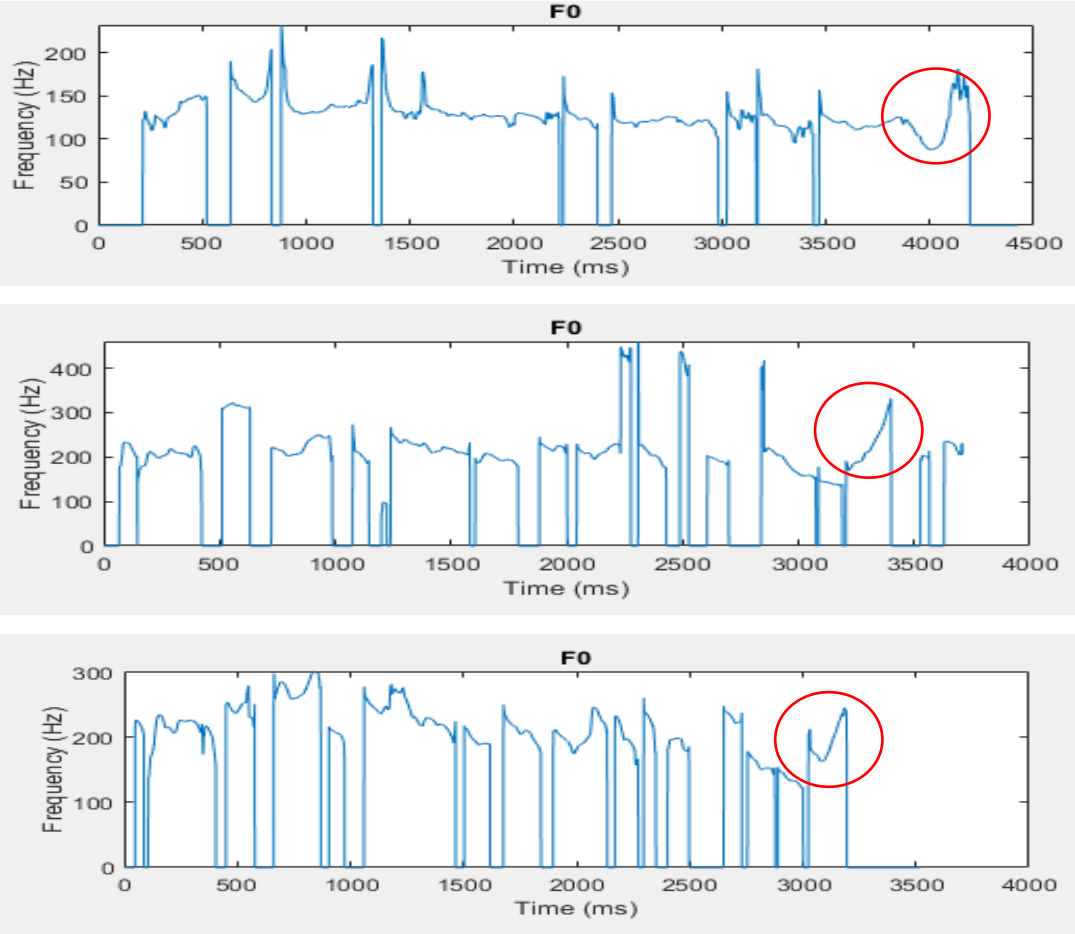
Kemudian akan diplot frekuensi dasar suara asli dan hasil sintesis suara *average voice model* untuk membandingkan frekuensi dasar suara asli dan hasil sintesis *average voice model*, identifikasi intonasi dan melihat kejelasan dari setiap kata yang disintesis. Pada Gambar 4.5 terdapat plot frekuensi dasar sintesis *average voice model* kalimat berita mmht, dan untuk Gambar 4.6 terdapat plot frekuensi dasar sintesis *average voice model* kalimat tanya mmht.







Gambar 4. 5 Plot frekuensi dasar sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training*

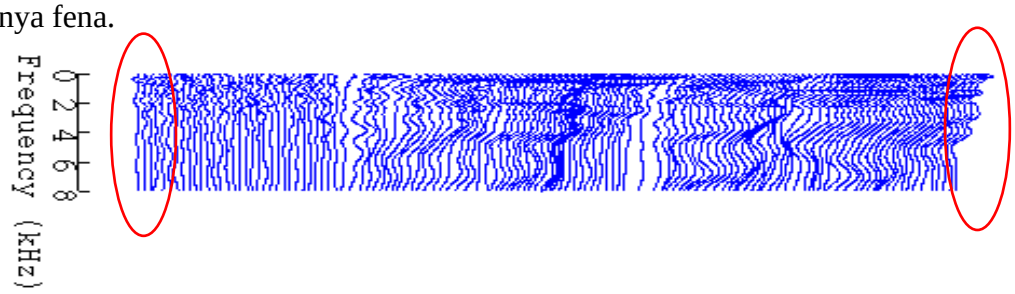


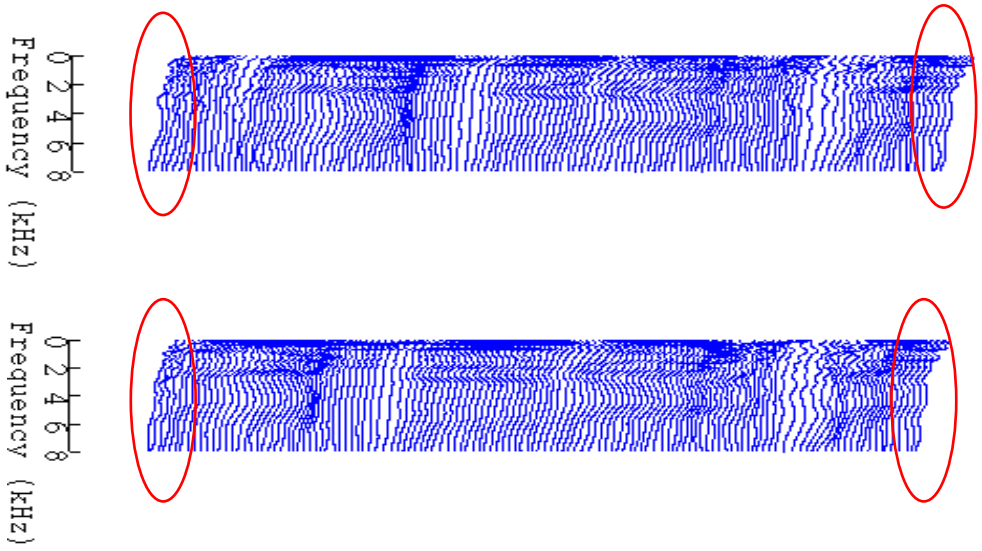
Gambar 4. 6 Plot frekuensi dasar sintesis suara mmht kalimat tanya "benarkah ayah akan meninggalkan kami untuk waktu yang lama ?" (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training*

Gambar 4.5 – 4.6 menunjukkan plot frekuensi dasar pembicara mmht dengan hasil sintesis suara *average voice model* kalimat berita dan kalimat tanya. Sumbu-x menunjukkan informasi waktu dan sumbu-y menunjukkan informasi nilai frekuensi dasar. Plot frekuensi dasar dilakukan menggunakan *software* matlab.

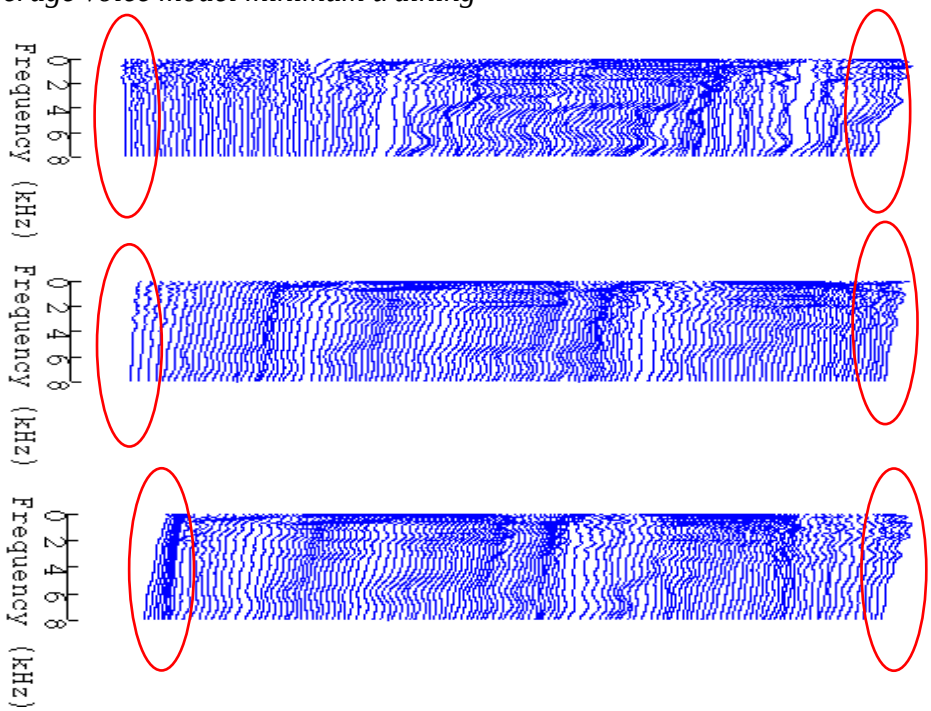
Sinyal suara disampling pada frekuensi sampling sebesar 16 kHz dan menggunakan 25 ms jendela *Blackman* dengan pergeserann 5 ms. Fitur vektor terdiri dari 25 *mel-cepstral coefficient*, log frekuensi dasar, dan delta-deltanya. Nilai frekuensi dasar menunjukkan nilai dimana getaran pita suara yang menghasilkan suara. Plot frekuensi dasar dapat menunjukkan informasi daerah *silence*, *voice*, dan *unvoiced* dari suara pengucapan suatu kalimat. Berdasarkan Gambar 4.5 – 4.6, dapat dilihat bahwa terdapat distorsi nilai frekuensi dasar dari hasil suara sintesis *average voice model* terhadap suara aslinya yang disebabkan karena nilai frekuensi dasar *average voice model* merupakan nilai frekuensi dasar rata-rata dari seluruh *speaker*. Kemudian, plot nilai frekuensi terlihat bergeser ke kiri mendahului plot nilai frekuensi suara aslinya dikarenakan durasi setiap speaker yang berbeda. Melalui lingkaran merah dapat dilihat perbedaan frekuensi dasar yang dihasilkan dari masing-masing kalimat sintesis, dimana pada kalimat tanya akan menghasilkan kontur F_0 yang naik di akhir kalimat dan akan menghasilkan kontur F_0 yang turun pada kalimat berita.

Selanjutnya pada Gambar 4.7 - 4.10 akan diplot spektral dari hasil sintesis *average voice model* untuk mengetahui bentuk spektral setiap fonem dan perpindahannya untuk menuju fonem selanjurnya. Pada Gambar 4.7 terdapat plot spektral sintesis *average voice model* kalimat berita mmht, untuk Gambar 4.8 terdapat plot spektral sintesis *average voice model* kalimat tanya mmht, untuk Gambar 4.9 terdapat plot spektral sintesis *average voice model* kalimat berita fena, dan untuk Gambar 4.10 terdapat plot spektral sintesis *average voice model* kalimat tanya fena.

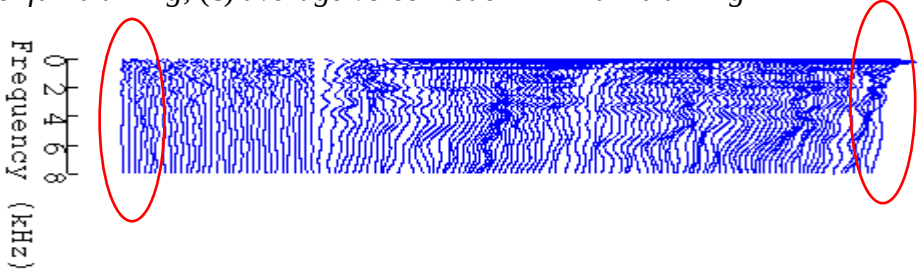


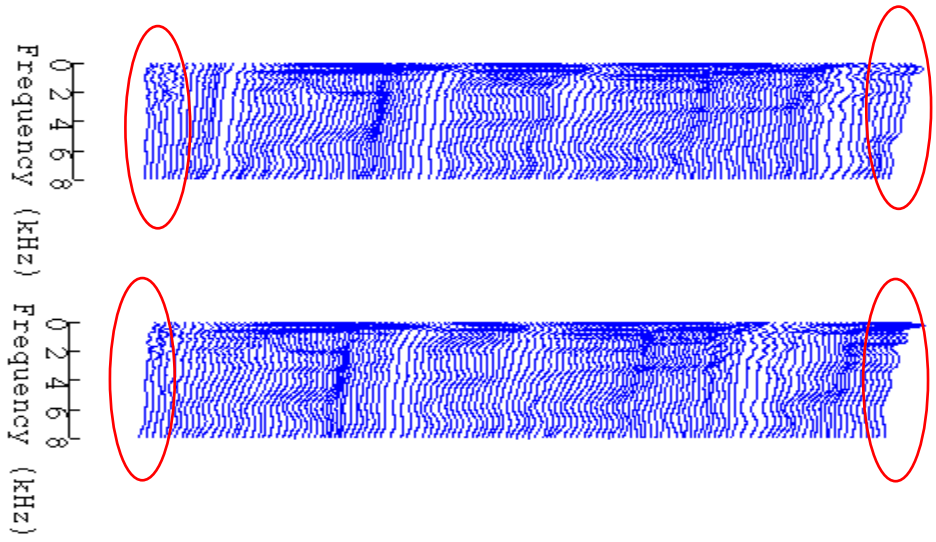


Gambar 4. 7 Plot spektral sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *average voice model full training*, (c)*average voice model minimum training*

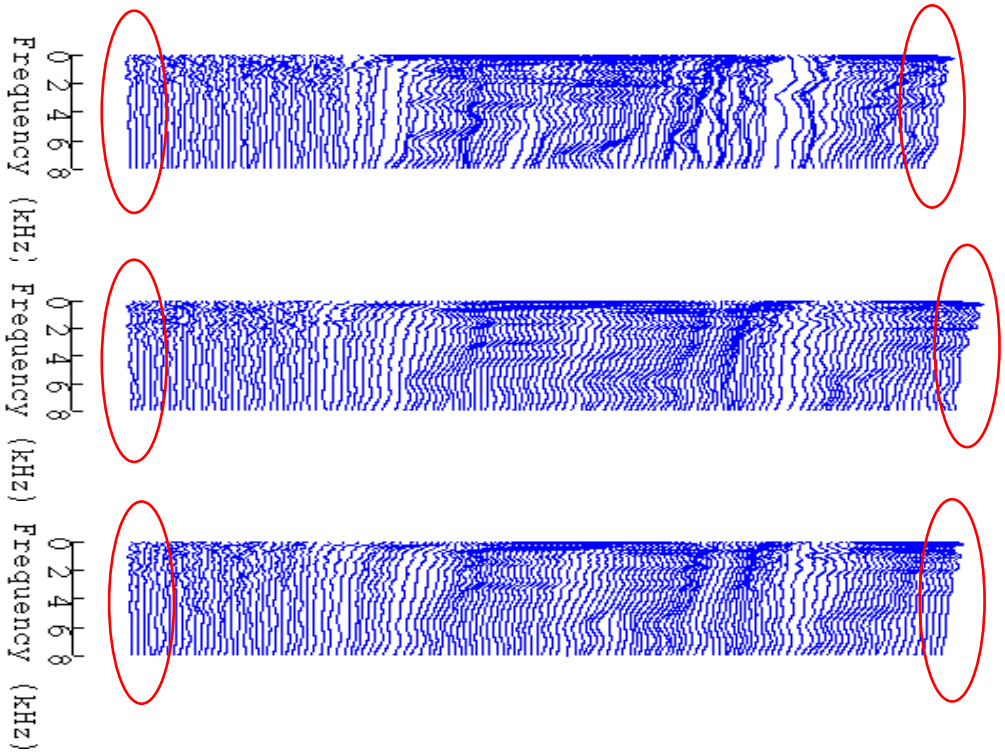


Gambar 4. 8 Plot spektral sintesis suara mmht kalimat tanya "benarkah ayah akan meninggalkan kami untuk waktu yang lama ?" (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training*





Gambar 4. 9 Plot spektral sintesis suara fena kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training*



Gambar 4. 10 Plot spektral sintesis suara fena kalimat tanya "benarkah ayah akan meninggalkan kami untuk waktu yang lama ?" (a) suara asli, (b) *average voice model full training*, (c) *average voice model minimum training*, (e) *speaker adaptation minimum training*

Plot spektral pada Gambar 4.7 – 4.10 menggambarkan spektral suara asli dan hasil sintesis *average voice model* yang sesuai dengan fonem pada kalimat tersebut. Sumbu-x pada grafik menunjukkan informasi fonem yang berubah terhadap waktu

dan sumbu-y menunjukkan informasi frekuensi. Dapat dilihat bahwa spektral pada sintesis *average voice* dan *minimum training* mengalami distorsi yang dapat dilihat pada lingkaran berwarna merah. Pada spektral *average voice* terdapat perbedaan di setiap fonemnya jika dibandingkan dengan spektral suara asli, hal ini dikarenakan pada *average voice* merupakan spektral gabungan semua pembicara pada tahap *training*. Kemudian dapat dilihat bahwa *silence* terjadi saat terdapat regangan antar spektral suara asli, namun pada hasil sintesis posisi *silence* berubah atau bahkan tidak ada. Sehingga, spektral *average voice* berbeda dengan spektral suara aslinya.

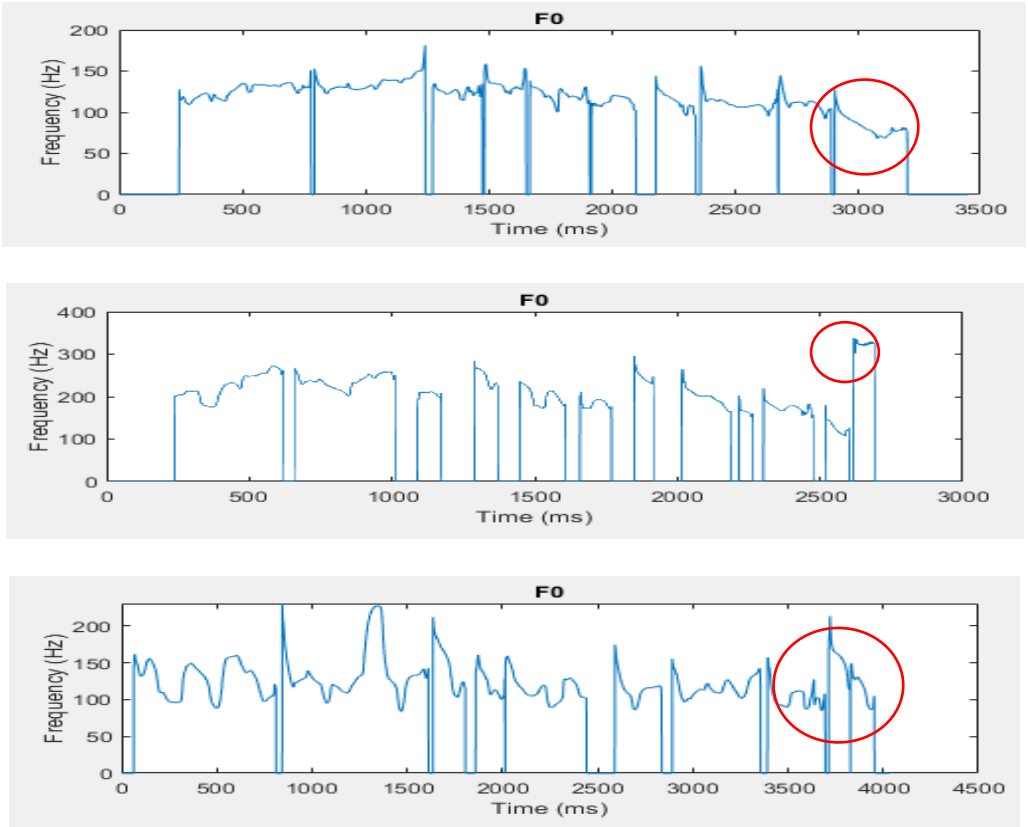
4.4 Hasil Sintesis *Speaker Adaptation* Bahasa Indonesia

Model akustik yang didapatkan dari tahap adaptasi akan menghasilkan kalimat sintesis dengan melewati MLSA filter untuk menjadi suara dengan format file ".wav". Identifikasi suara sintesis dapat dilihat pada plot *waveform,* frekuensi dasar, dan spektral yang menunjukkan intonasi dari kalimat, kejelasan pada setiap kata, dan susunan antar fonem pada kalimat. *Waveform* untuk sintesis suara *speaker adaptation* mmht kalimat berita terdapat pada Gambar 4.11, Plot *waveform* dihasilkan menggunakan *software* wavesurfer

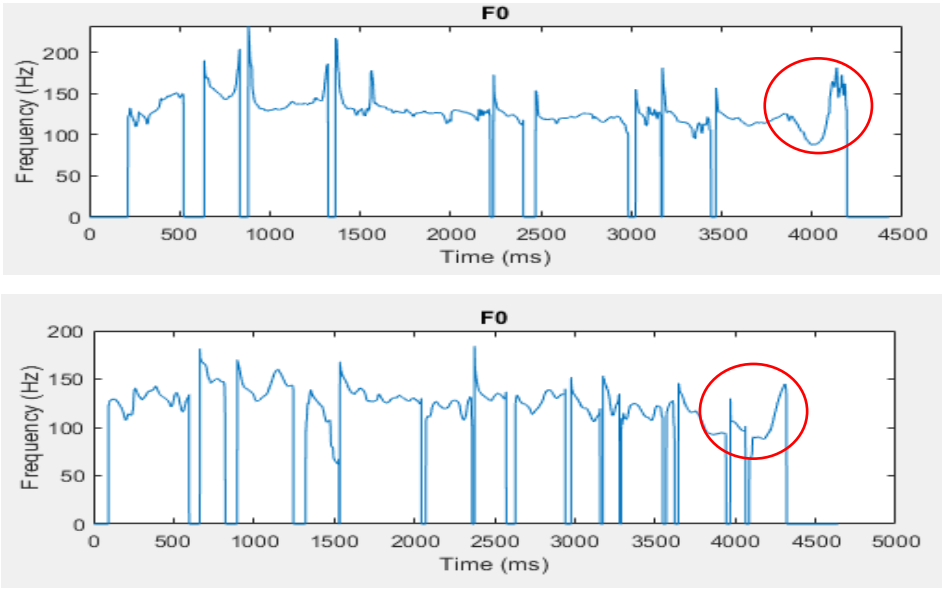
Gambar 4. 11 Plot *waveform* sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *speaker adaptation full training,* (c) *speaker adaptation minimum training*

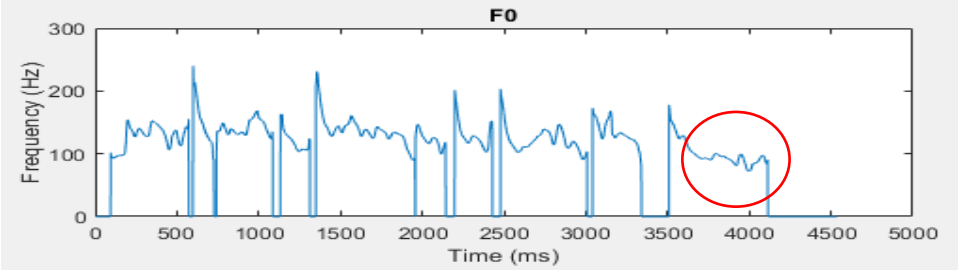
Pada Gambar 4.11 merupakan plot perbandingan *waveform* suara asli dan suara hasil sintesis *speaker adaptation* kalimat berita pada pembicara mmht. Sumbu-x menunjukkan informasi waktu dan sumbu-y menunjukkan informasi amplitudo sinyal suara. Pada gambar tersebut terlihat jelas terdapat perbesaran amplitudo pada *waveform* hasil sintesis yang disebabkan adanya amplifikasi sehingga terjadi clipping yang menyebabkan suara terdengan seperti terpotong-potong dan juga terdapat perbedaan durasi suara yang diakibatkan oleh tidak terbacanya *unvoiced* setiap kalimat, sehingga tidak adanya *silence* atau jeda saat awal kalimat, akhir kalimat dan antar kata setiap kalimat. Oleh karena itu, suara dari hasil sintesis *speaker adaptation* jelas terdengar setiap katanya namun terdapat perbedaan durasi saat memulai kalimat dan durasi dari masing-masing kalimat itu sendiri.

Kemudian akan diplot frekuensi dasar suara asli dan hasil sintesis suara untuk membandingkan frekuensi dasar suara asli dan hasil sintesis *speaker adaptation*, identifikasi intonasi dan melihat kejelasan dari setiap kata yang disintesis. Jika diperhatikan dari kontur F_0, sintesis kalimat tanya dan berita memiliki beberapa perbedaan. Pada kalimat tanya memiliki ciri intonasi yang naik pada akhir kalimat. Sedangkan kalimat berita memiliki ciri intonasi yang datar pada akhir kalimat. Pada Gambar 4.12 terdapat plot frekuensi dasar sintesis *speaker adaptation* kalimat berita mmht, dan untuk Gambar 4.13 terdapat plot frekuensi dasar sintesis *speaker adaptation* kalimat tanya mmht. Plot frekuensi dasar dilakukan menggunakan *software* matlab. Sinyal suara disampling pada frekuensi sampling sebesar 16 kHz dan menggunakan 25 ms jendela *Blackman* dengan pergeserann 5 ms. Fitur vektor terdiri dari 25 *mel-cepstral coefficient*, log frekuensi dasar, dan delta-deltanya.



Gambar 4. 12 Plot frekuensi dasar sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *speaker adaptation full training*, (c)*speaker adaptation minimum training*



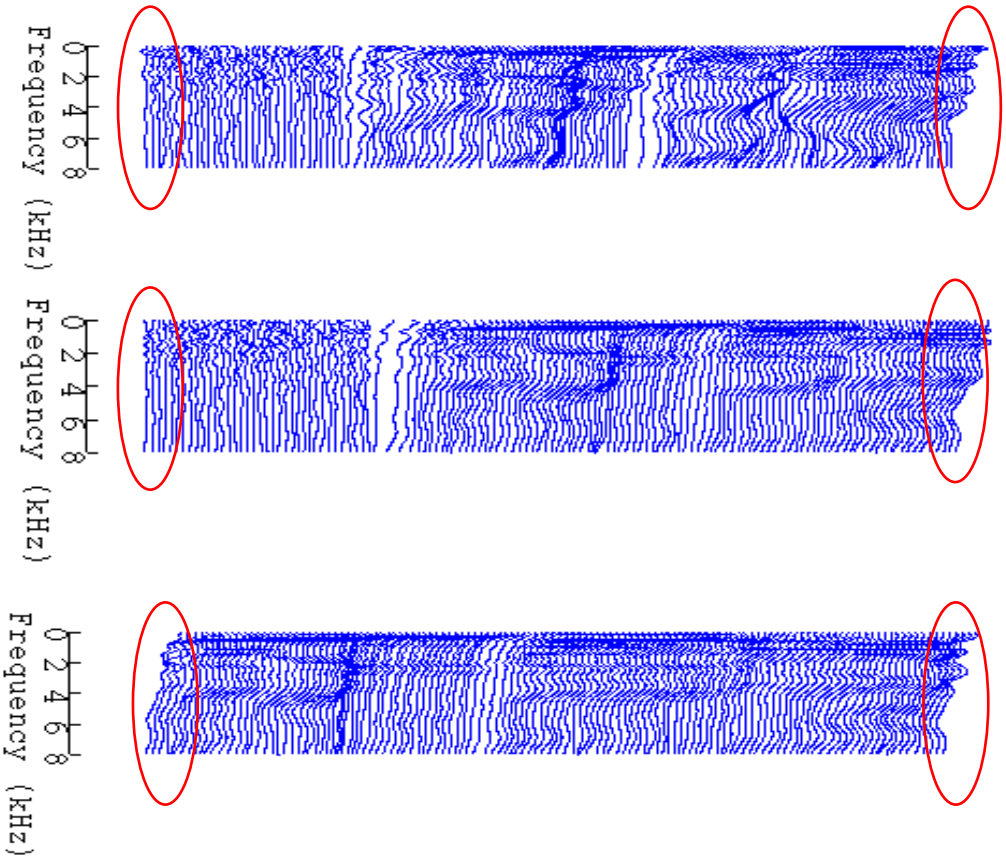


Gambar 4. 13 Plot frekuensi dasar sintesis suara mmht kalimat tanya "benarkah ayah akan meninggalkan kami untuk waktu yang lama ?" (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*

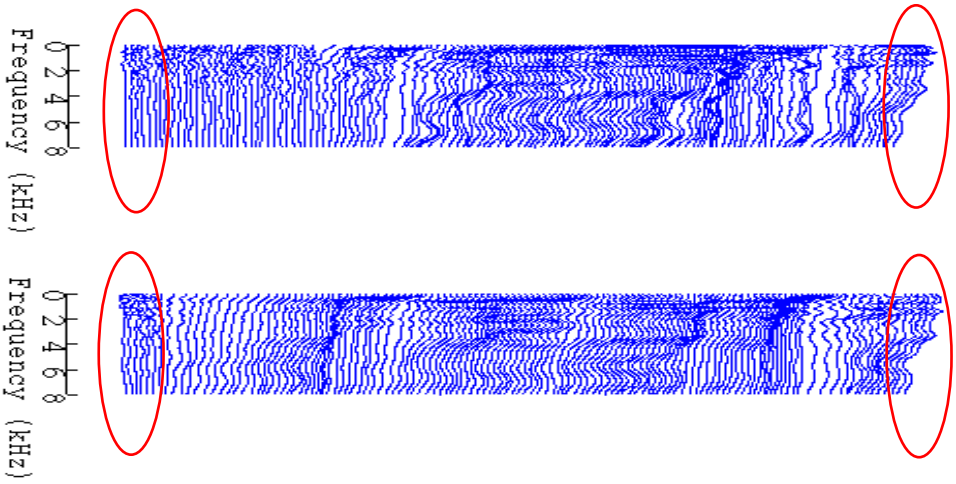
Gambar 4.12 – 4.13 menunjukkan plot frekuensi dasar pembicara mmht dengan hasil sintesis *speaker adaptation* suara kalimat berita dan kalimat tanya. Sumbu-x menunjukkan informasi waktu dan sumbu-y menunjukkan informasi nilai frekuensi dasar. Nilai frekuensi dasar menunjukkan nilai dimana getaran pita suara yang menghasilkan suara. Plot frekuensi dasar dapat menunjukkan informasi daerah *silence*, *voice*, dan *unvoiced* dari suara pengucapan suatu kalimat. Berdasarkan Gambar 4.12 – 4.13, dapat dilihat bahwa terdapat distorsi nilai frekuensi dasar dari hasil suara sintesis *speaker adaptation* terhadap suara aslinya. Pada hasil sintesis suara *speaker adaptation*, plot nilai frekuensi terlihat bergeser ke kiri mendahului plot nilai frekuensi suara aslinya, hal ini menunjukkan sinyal suara bergerak lebih dahulu yang menyebabkan daerah *silence*, *voice*, dan *unvoiced* juga ikut bergeser ke kiri atau bahkan tidak terbaca atau tidak ada. Adanya pergeseran suara terjadi akibat tidak terbacanya fonem *silence* pada saat sintesis suara. Hal ini menyebabkan suara terdengar putus-putus. Kemudian, melalui lingkaran merah dapat dilihat perbedaan frekuensi dasar yang dihasilkan dari masing-masing kalimat sintesis, dimana pada kalimat tanya akan menghasilkan kontur F_0 yang naik di akhir kalimat dan akan menghasilkan kontur F_0 yang turun pada kalimat berita.

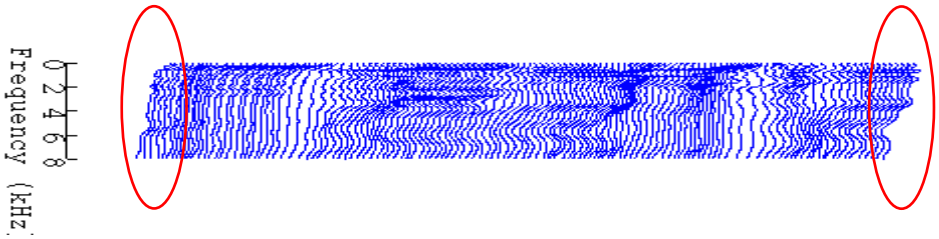
Selanjutnya pada Gambar 4.14 - 4.17 akan diplot spektral dari hasil sintesis untuk mengetahui bentuk spektral setiap fonem dan perpindahannya untuk menuju fonem selanjurnya. Pada Gambar 4.14 terdapat plot spektral sintesis *speaker adaptation* kalimat berita mmht, untuk Gambar 4.15 terdapat plot spektral sintesis *speaker adaptation* kalimat tanya mmht, untuk Gambar 4.16 terdapat plot spektral

sintesis *speaker adaptation* kalimat berita fena, dan untuk Gambar 4.17 terdapat plot spektral sintesis *speaker adaptation* kalimat tanya fena.

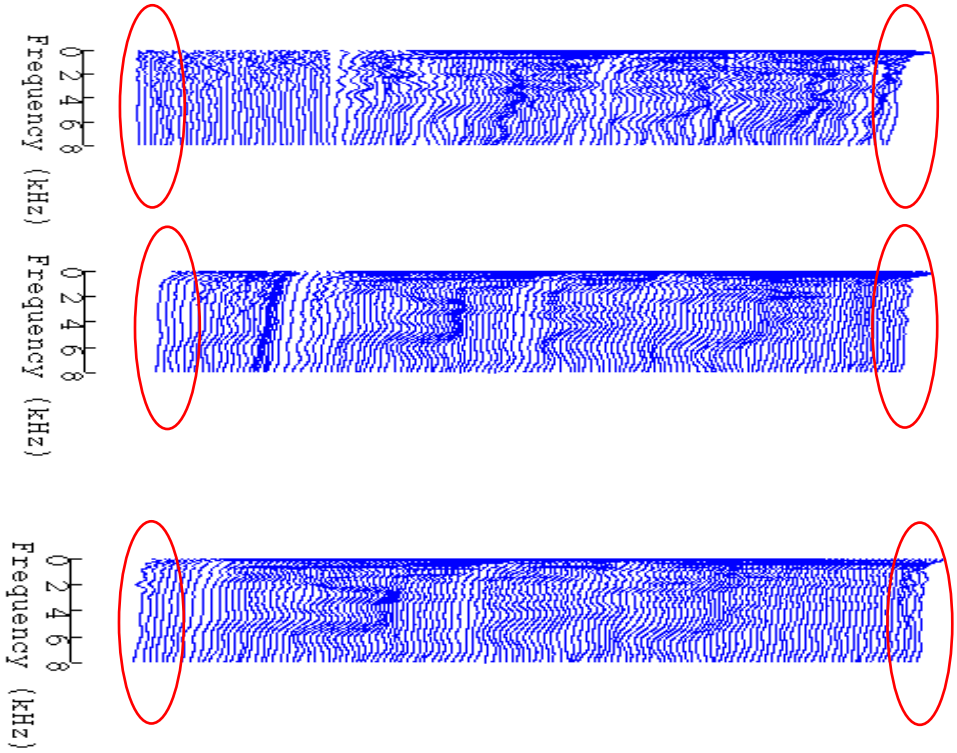


Gambar 4. 14 Plot spektral sintesis suara mmht kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*

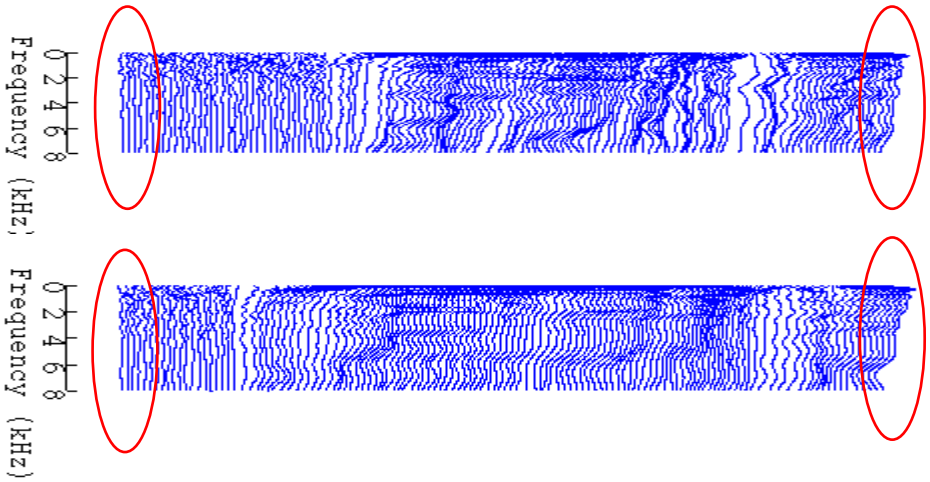


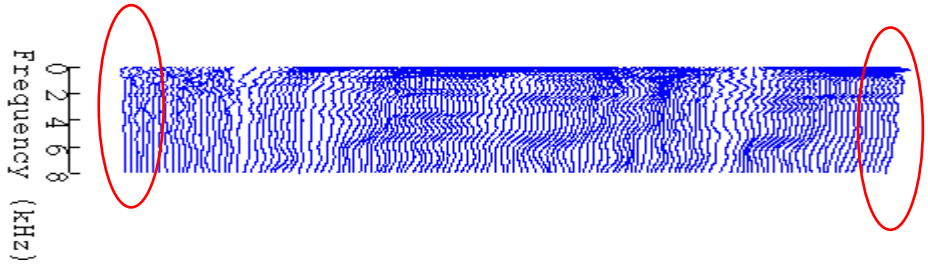


Gambar 4. 15 Plot spektral sintesis suara mmht kalimat tanya "benarkah ayah akan meninggalkan kami untuk waktu yang lama ?" (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*



Gambar 4. 16 Plot spektral sintesis suara fena kalimat berita "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*



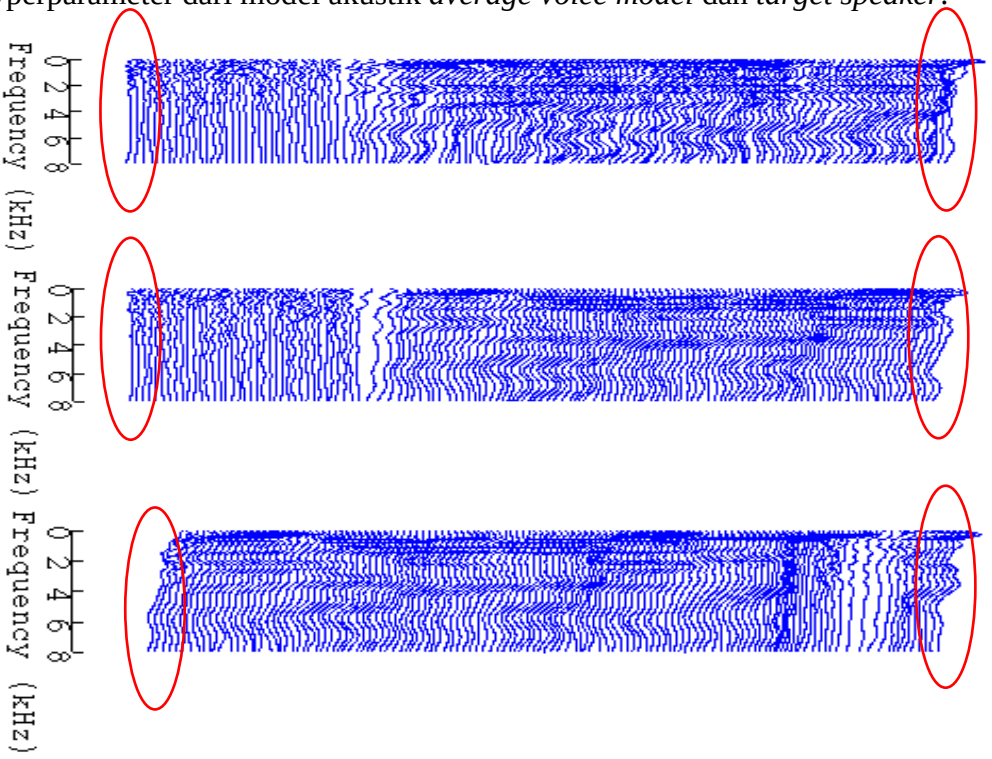


Gambar 4. 17 Plot spektral sintesis suara fena kalimat tanya "benarkah ayah akan meninggalkan kami untuk waktu yang lama ?" (a) suara asli, (b) *average voice model full training*, (c) *speaker adaptation full training*, (c) *speaker adaptation minimum training*

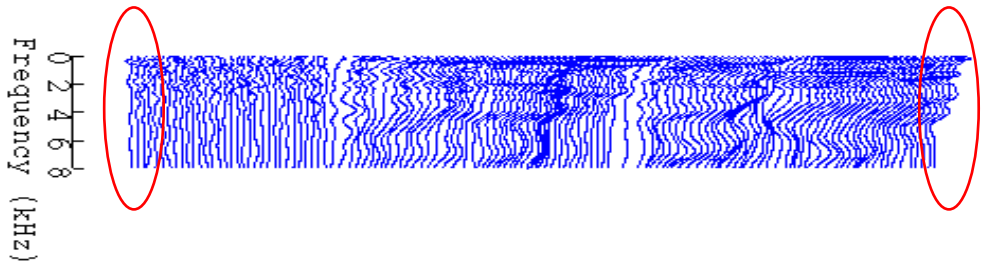
Plot spektral pada Gambar 4.14 – 4.17 menggambarkan spektral suara asli dan hasil sintesis *speaker adaptation* yang sesuai dengan fonem pada kalimat tersebut. Sumbu-x pada grafik menunjukkan informasi fonem yang berubah terhadap waktu dan sumbu-y menunjukkan informasi frekuensi. Dapat dilihat bahwa spektral pada sintesis *speaker adaptation full training* sangat mirip dengan spektral suara aslinya, sedangkan pada spektral hasil sintesis *minimum training* mengalami distorsi yang dapat dilihat pada lingkaran berwarna merah. Pada spektral dengan *minimum training* terlihat terjadi *oversmoothing* pada akhir kalimat yang menyebabkan hilangnya efek *global variance* dan adanya perubahan spektral pada setiap frekuensinya. Kemudian dapat dilihat bahwa *silence* terjadi saat terdapat regangan antar spektral suara asli, namun pada hasil sintesis posisi *silence* berubah atau bahkan tidak ada. Sehingga, spektral *minimum training* berbeda dengan spektral suara aslinya.

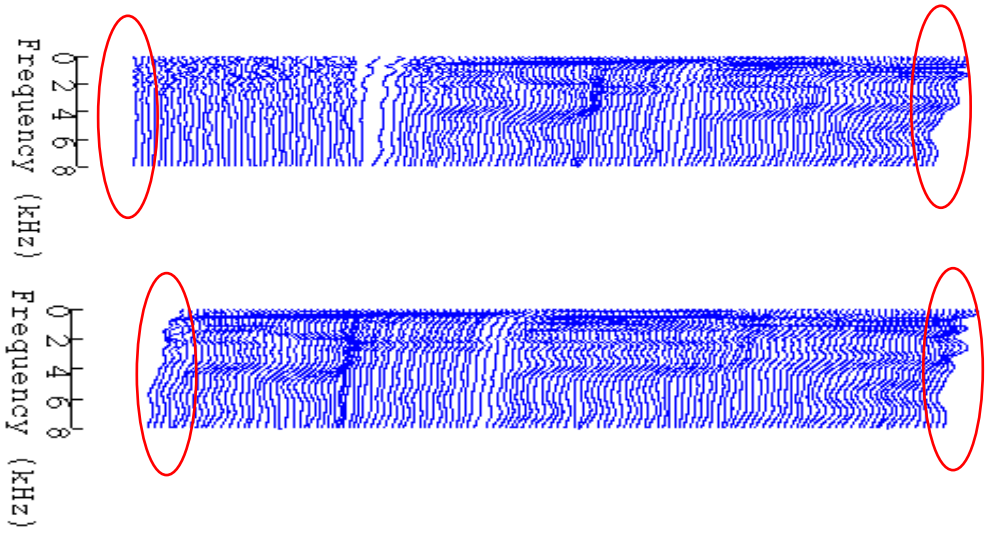
Jika diperhatikan pada plot *waveform* dan frekuensi dasar, hasil sintesis suara *speaker adaptation* belum persis sama dengan suara asli. Hal ini dikarenakan proses transformasi menggunakan MLLR dan CMLLR sudah berjalan dengan baik, namun terdapat kekurangan basis data pada target speaker yang akan menjadi estimasi dari model yang akan diadaptasi. Pada CMLLR terdapat regresi linier antara parameter mean dan kovarian, sehingga proses transformasi atau adaptasi dapat lebih akurat dilakukan dengan pendekatan kepada fitur akustik, sehingga pada CMLLR membutuhkan ketepatan untuk estimasi fitur akustik yang membutuhkan durasi suara lebih dari 5 detik untuk adaptasinya (Povey & Yao, 2012). Pada sistem ini memiliki batasan durasi basis data pada tahap adaptasi, sehingan hanya 30 basis data yang digunakan untuk membentuk model parameter akustik pada tahap

adaptasi dari 1529 basis data yang diinput sehingga mempengaruhi kualitas hasil sintesis. Jika dibandingkan dengan jurnal (Yamagishi, Kobayashi, Nakano, Ogata, & Isogai, 2009) yang menggunakan 100 basis data pada tahap adaptasi akan menghasilkan transformasi yang lebih baik. Pada jurnal tersebut juga membandingkan metode CMLLR dan CSMAPLR, dimana metode CSMAPLR lebih baik dikarenakan estimasi dan transformasi adaptasi hingga ke hyperparameter dari model akustik *average voice model* dan *target speaker*.



Gambar 4. 18 Plot spektral perbandingan suara mmht kalimat “malam itu paman menonton televisi di kamar” (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*





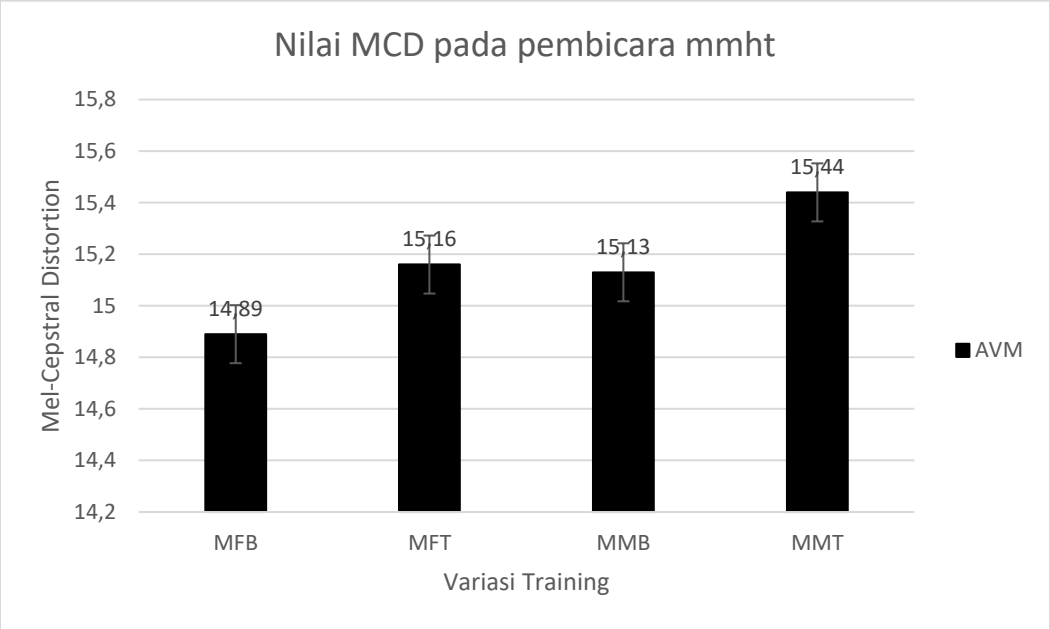
Gambar 4. 19 Plot spektral perbandingan suara mmht kalimat "liburan kemarin aku tidak bisa pulang kampung" (a) suara asli, (b) *speaker adaptation full training*, (c) *speaker adaptation minimum training*

Pada Gambar 4.18 dan 4.19 terdapat plot spektral mmht kalimat berita "malam itu paman menonton televisi di kamar" dan "liburan kemarin aku tidak bisa pulang kampung" dengan perbandingan suara asli, *speaker adaptation full training*, dan *speaker adaptation minimum training*. Jika dilihat dari jahitan spektralnya terhadap fonem pada kalimat, hasil spektral variasi *full training* mendekati spektral suara asli dibandingkan dengan *minimum training*. Kemudian jika dibandingkan Gambar 4.18 dan 4.19, hasil speaker adaptation pada kalimat "malam itu paman menonton televisi di kamar" lebih baik dari kalimat "liburan kemarin aku tidak bisa pulang kampung", hal ini dikarenakan dari lebih banyaknya kata dan fonem yang dilatih pada kalimat "malam itu paman menonton televisi di kamar" daripada kalimat "liburan kemarin aku tidak bisa pulang kampung".

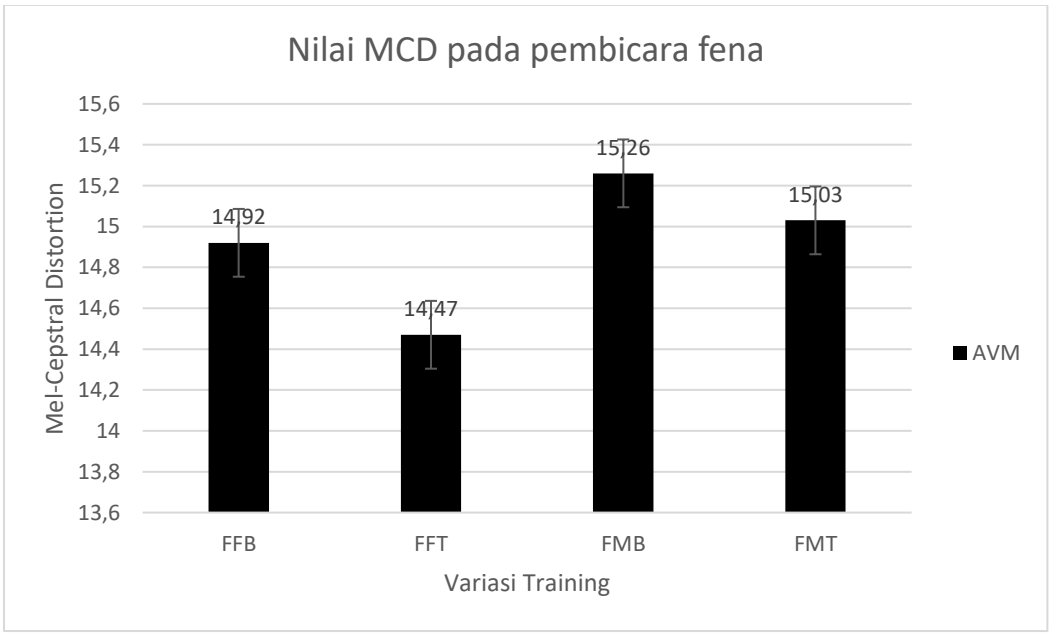
4.5 Hasil Pengujian Objektif

Pengujian objektif dilakukan dengan menggunakan metode *mel-cepstrum distance* (MCD) dan *root mean square error* (RMSE) pada nilai log frekuensi dasar. Semakin kecil nilai MCD menunjukkan kecilnya perbedaan antara log frekuensi dasar pada suara sintesis yang dihasilkan dengan log frekuensi dasar suara aslinya sebagai *baseline*. Sehingga semakin kecil nilai MCD maka akan semakin natural suara yang disintesis dan begitupula sebaliknya. Nilai persentase RMSE merepresentasikan nilai error dari frekuensi dasar yang disintesis, semakin kecil

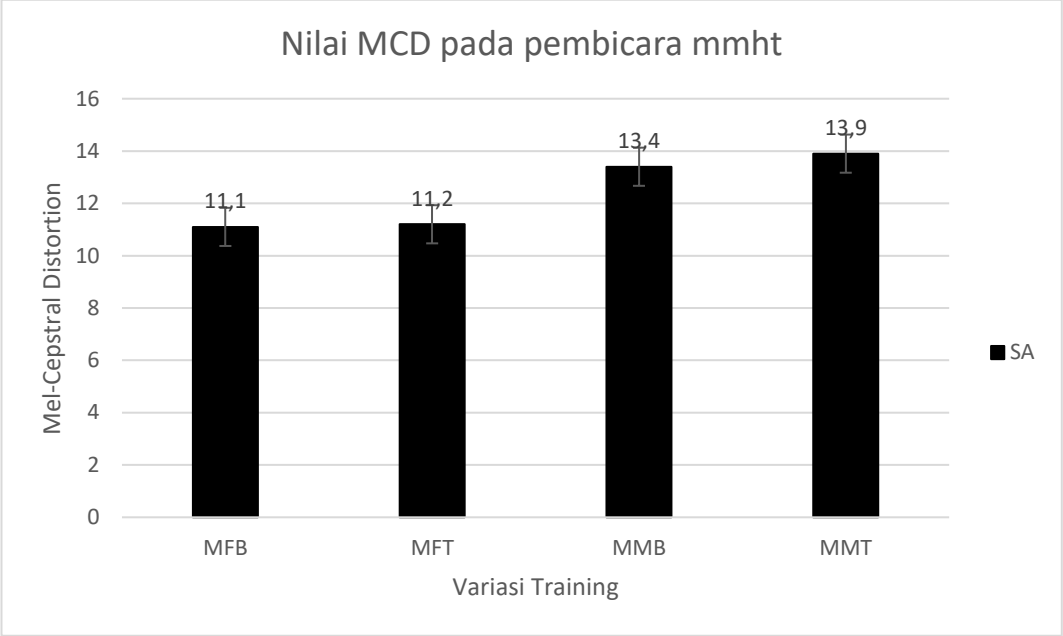
nilai RMSE maka suara sintesis semakin mendekati suara aslinya. Pengujian dilakukan menggunakan *tools* '*cdist*'untuk mencari nilai MCD pada SPTK dan *tools* '*rmse*' untuk mencari nilai RMSE pada SPTK.



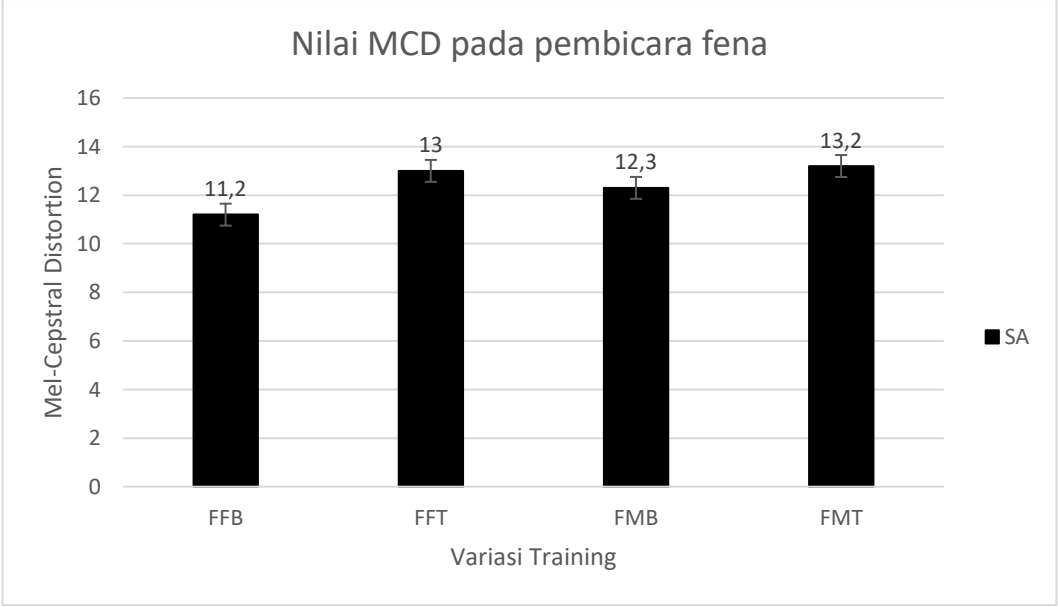
Gambar 4. 20 Hasil pengujian objektif MCD *average voice model* pada pembicara mmht



Gambar 4. 21 Hasil pengujian objektif MCD *average voice model* pada pembicara fena



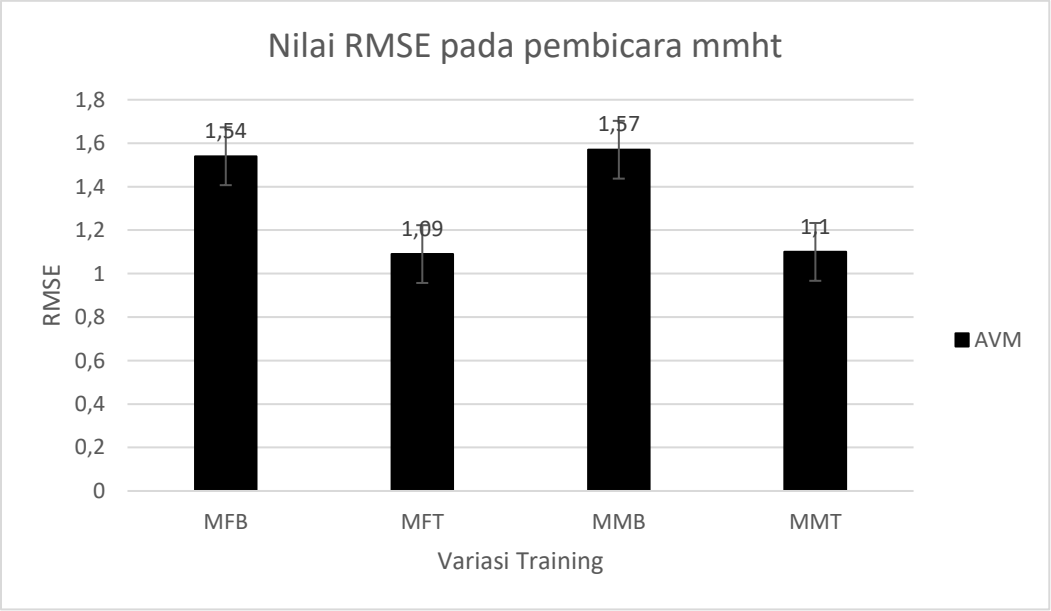
Gambar 4. 22 Hasil pengujian objektif MCD *speaker adaptation* pada pembicara mmht



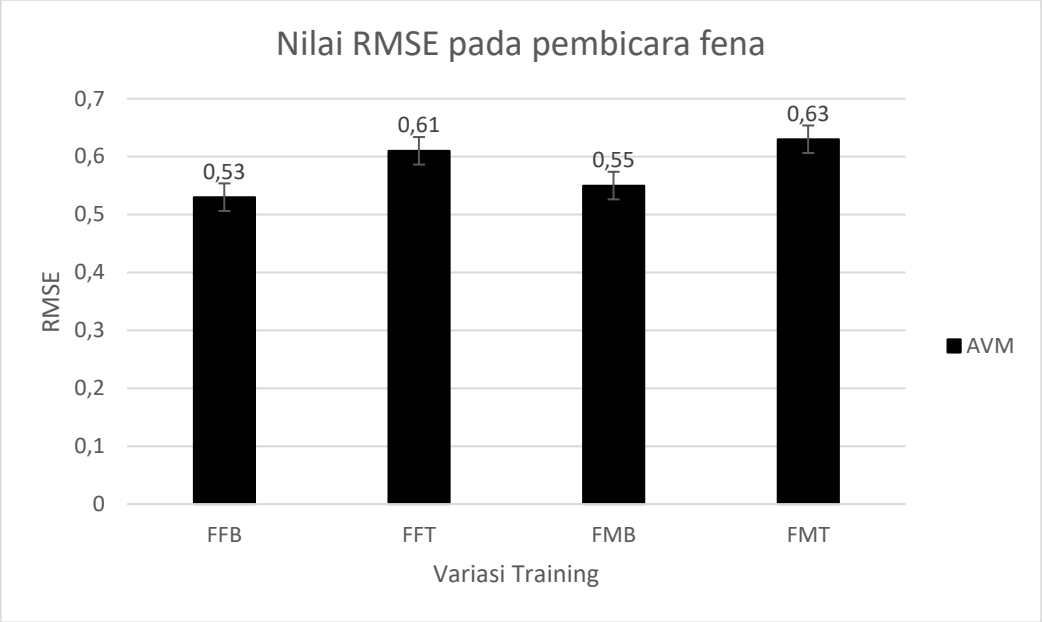
Gambar 4. 23 Hasil pengujian objektif MCD *speaker adaptation* pada pembicara fena

Hasil pengujian objektif MCD ditunjukkan oleh Gambar 4.20-4.21 untuk pembicara mmht dan fena dengan metode *average voice model* dan Gambar 4.22-4.23 untuk pembicara mmht dan fena dengan metode *speaker adaptation* dengan kalimat tanya dan kalimat berita. Pengujian dilakukan menggunakan *tools* '*cdist*'untuk mencari nilai MCD pada SPTK. Informasi variasi training disingkat

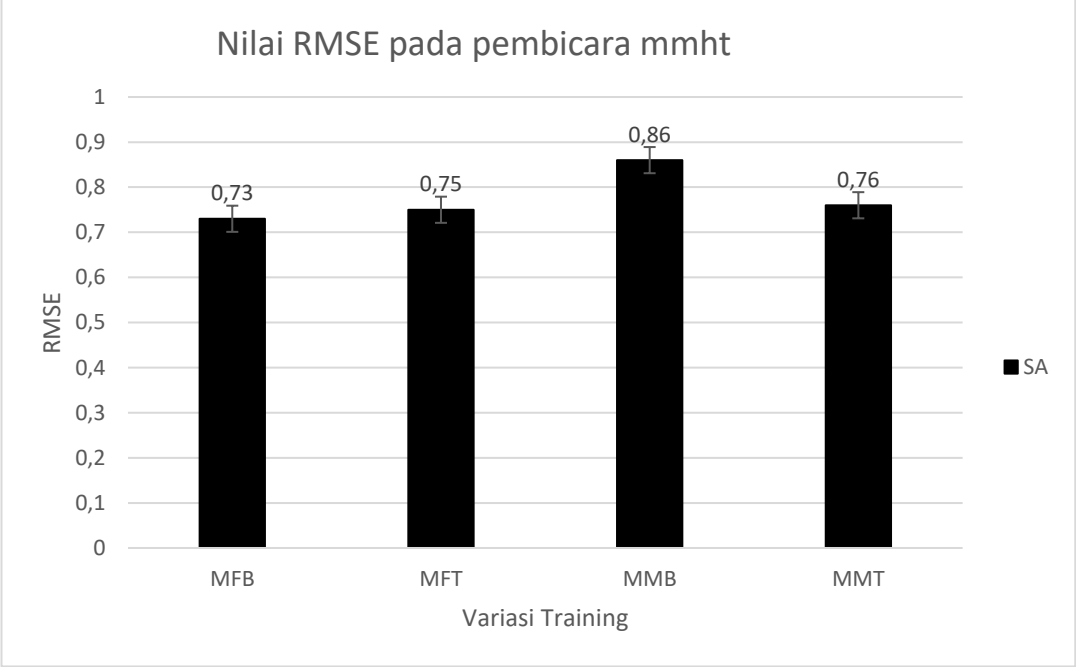
menjadi MFB (mmht *full training* kalimat berita), MFT (mmht *full training* kalimat tanya), MMB (mmht *minimum taining* kalimat berita), MMT (mmht *minimum training* kalimat tanya), FFB (fena *full training* kalimat berita), FFT (fena *full training* kalimat tanya), FMB (fena *minimum training* kalimat berita), dan FMT (fena *minimum training* kalimat tanya). Sumbu-x merupakan informasi variasi *training* yang digunakan saat sintesis, sumbu-y merupakan informasi nilai MCD. Berdasarkan Gambar 4.20-4.21 menunjukkan bahwa kualitas suara sintesis masih buruk pada metode *average voice model* dengan diperolehnya nilai MCD kurang lebih 15 dan pada Gambar 4.22-4.23 menunjukkan bahwa kualitas suara sintesis cukup baik dengan nilai MCD mendekati nilai 10. Nilai distorsi paling kecil pada suara sintesis fena kalimat berita *full training* dengan nilai 11,2 dan pada suara sintesis mmht 11,1 pada kalimat berita *full training*. Pada Gambar 4.22 dan 4.23, dapat diamati bahwa nilai MCD pada *speaker adaptation* lebih baik daripada nulai *MCD average voice model*. Berdasarkan hasil analisa MCD, distorsi *mel-cepstral* akan semakin kecil pada hasil sintesis *full training*, hal ini menandakan bahwa semakin banyak basis data maka akan semakin baik kualitas sintesis suara yang dihasilkan karena banyaknya intensitas kemunculan fonem pada proses *training*.



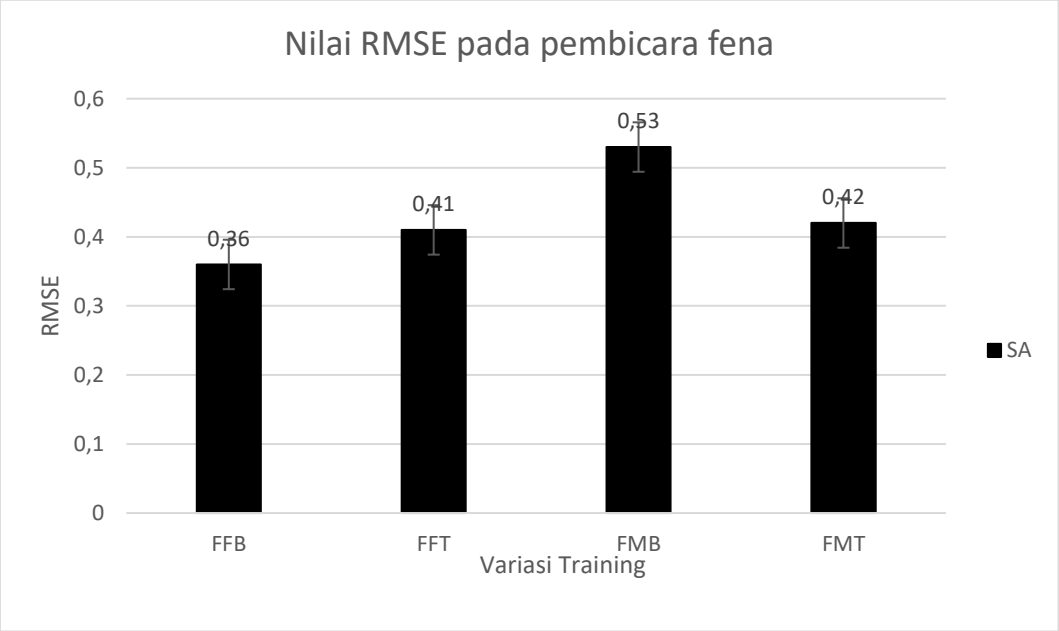
Gambar 4. 24 Hasil Pengujian objektif RMSE *average voice model* pada pembicara mmht



Gambar 4. 25 Hasil pengujian objektif RMSE *average voice model* pada pembicara fena



Gambar 4. 26 Hasil pengujian objektif RMSE speaker adaptation pada pembicara mmht



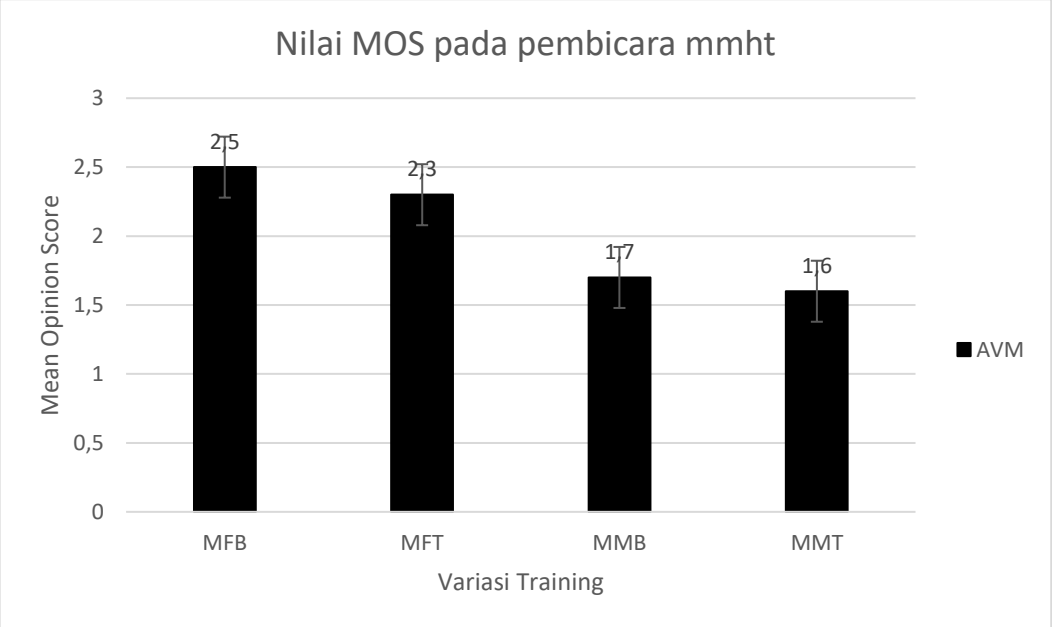
Gambar 4. 27 Hasil pengujian objektif RMSE *speaker adaptation* pada pembicara fena

Pengujian objektif dengan metode *root mean square error* (RMSE) dilakukan untuk mengukur perbedaan nilai log F_0 suara hasil sintesis dengan suara asli sebagai *baseline*. Pengujian dilakukan menggunakan *tools* '*rmse*' untuk mencari nilai RMSE pada SPTK. Nilai RMSE yang semakin kecil menunjukkan perbedaan besar perubahan frekuensi dasar semakin kecil sehingga memiliki kualitas suara yang lebih baik dan memiliki nilai F_0 mendekati atau sama dengan suara aslinya. Informasi variasi *training* pada Gambar 4.24-4.27, variasi *training* disingkat menjadi MFB (mmht *full training* kalimat berita), MFT (mmht *full training* kalimat tanya), MMB (mmht *minimum taining* kalimat berita), MMT (mmht *minimum training* kalimat tanya), FFB (fena *full training* kalimat berita), FFT (fena *full training* kalimat tanya), FMB (fena *minimum training* kalimat berita), dan FMT (fena *minimum training* kalimat tanya). Gambar 4.24-4.25 menunjukkan hasil uji objektif RMSE pembicara mmht dan fena dengan metode *average voice model* dan Gambar 4.26-4.27 menunjukkan hasil uji objektif RMSE pada pembicara mmht dan fena dengan metode *speaker adaptation*. Nilai RMSE paling rendah untuk pembicara mmht diperoleh oleh sintesi kalimat berita *full training* sebesar 0,73 dan nilai RMSE paling rendah untuk pembicara fena diperoeh oleh sintesis kalimat

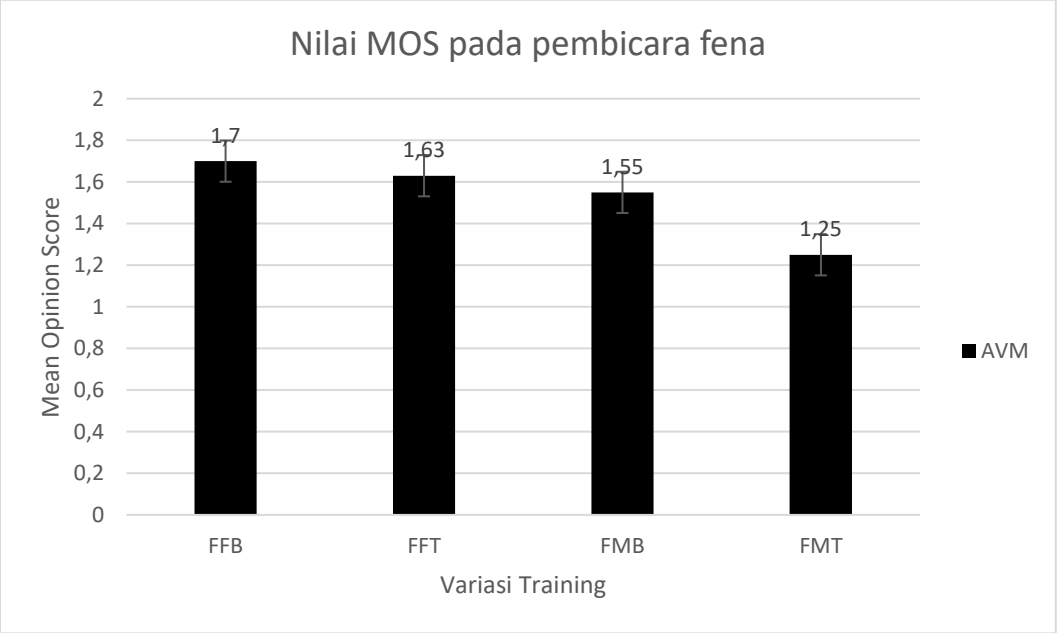
berita *full training* sebesar 0,36. Pada Gambar 4.24 dan 4.25, dapat diamati bahwa distorsi frekuensi dasar pada teknik *average voice model* lebih besar dibandingkan dengan teknik *speaker adaptation*, sehingga dapat disimpulkan bahwa teknik *speaker adaptation* menghasilkan sinstesis suara lebih baik dan lebih mirip dengan suara aslinya dibandingkan teknik *average voice model*. Berdasarkan hasil yang didapatkan, dengan jumlah basis data *training* yang banyak akan memberikan kualitas suara yang lebih baik dan distorsi frekuensi dasar yang lebih kecil.

4.6 Hasil Pengujian Subjektif

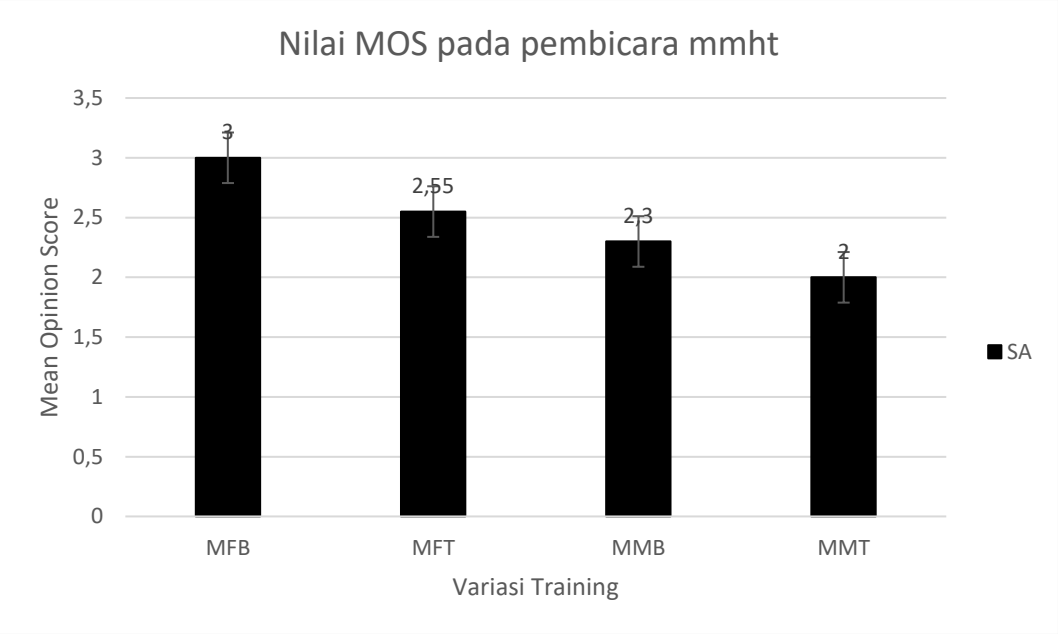
Pengujian subjektif dilakukan untuk menilai dan analisa kenaturalan suara hasil sintesis. Suara yang natural menunjukkan suara hasil sintesis mirip dengan suara aslinya dan sesuai dengan intonasi atau emosi kalimatnya. Pengujian dilakukan menggunakan metode *mean opinion score* (MOS), dimana naracoba yang mendengarkan secara acak suara asli dan suara hasil sintesis, kemudian naracoba melakukan penilaian terhadap masing-masing suara sesuai dengan kategori pada Tabel 2.2. Penilaian dilakukan oleh 16 orang naracoba dengan rentang usia 19-24 tahun yang dipilih secara acak.



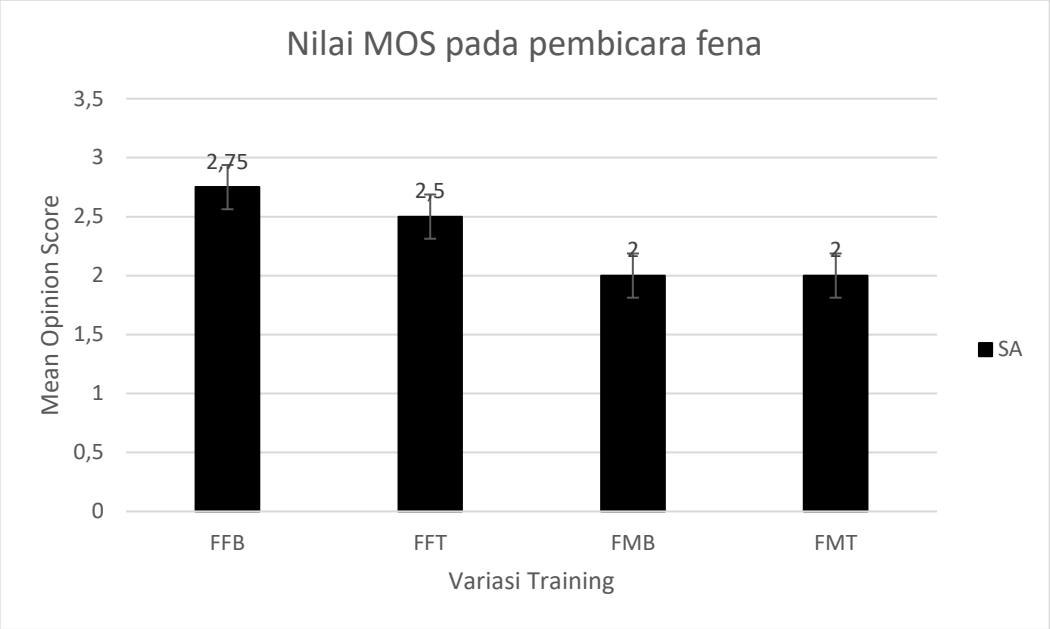
Gambar 4. 28 Hasil Pengujian subjektif metode *average voice model* pada pembicara mmht



Gambar 4. 29 Hasil pengujian subjektif metode *average voice model* pada pembicara fena



Gambar 4. 30 Hasil pengujian subjektif metode *speaker adaptation* pada pembicara mmht



Gambar 4. 31 Hasil pengujian subjektif metode *speaker adaptation* pada pembicara fena

Pada Gambar 4.28-4.31, informasi variasi *training* disingkat menjadi MFB (mmht *full training* kalimat berita), MFT (mmht *full training* kalimat tanya), MMB (mmht *minimum taining* kalimat berita), MMT (mmht *minimum training* kalimat tanya), FFB (fena *full training* kalimat berita), FFT (fena *full training* kalimat tanya), FMB (fena *minimum training* kalimat berita), dan FMT (fena *minimum training* kalimat tanya). Gambar 4.28-4.29 menunjukkan hasil uji subjektif sintesis suara pada pembicara mmht dan fena menggunakan metode *average voice model* dan pada Gambar 4.30-4.31 menunjukkan hasil uji subjektif sintesis suara pada pembicara mmht dan fena menggunakan metode *speaker adaptation*, dengan sumbu-x merupakan variasi *training* dan sumbu-y merupakan nilai MOS. Nilai MOS dikategorikan menjadi nilai 1-5, dengan semakin besar nya nilai maka semakin natural nya suara. Gambar 4.30-4.31 menunjukkan bahwa sintesis suara terbaik pada teknik *speaker adaptation* pembicara mmht kalimat berita *full training* 2,75/5 dan pada pembicara fena kalimat berita *full training* 3/5. Nilai tersebut menunjukkan bahwa kualitas sintesis suara cukup baik. Berdasarkan hasil uji objektif dan subjektif, dapat disimpulkan hasil sintesis *speaker adaptation* lebih baik dibandingkan dengan *average voice model*.

Halaman ini sengaja dikosongkan

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, didapatkan kesimpulan sebagai berikut:

a. Diperoleh karakteristik suara sintesis *speaker adaptation* dalam bahasa indonesia pada kalimat berita dan kalimat tanya dengan jumlah basis data *full training* 1529 kalimat dan *minimum training* 120 kalimat. Terdapat 8 variasi hasil sintesis, pembicara laki-laki (mmht) dan perempuan (fena) dengan hasil sintesis berupa kalimat tanya dan kalimat berita dengan variasi *full training* dan *minimum training*.
b. Diperoleh nilai kualitas suara sintesis *speaker adaptation* lebih baik dibandingkan dengan kualitas suara *average voice model*. Berdasarkan pengujian objektif dengan metode MCD dan RMSE. Nilai MCD terbaik pembicara fena sebesar 11,2 pada *full training* kalimat berita dan pada pembicara mmht sebesar 11,1 pada *full training* kalimat berita. Nilai RMSE terbaik pembicara fena sebesar 0,73 pada *full training* kalimat tanya dan pembicara mmht sebesar 0,36 pada *full training* kalimat tanya. Berdasarkan uji subjektif, nilai MOS pada pembicara fena kalimat berita *full training* sebesar 2,75/5 dan pada pembicara mmht kalimat berita *full training* sebesar 3/5. Sehingga hasil sintesis suara dapat digolongkan "cukup baik".

5.2 Saran

Pada penelitian selanjutnya, sebaiknya dilakukan perbaikan untuk meningkatkan unjuk kerja pada teknik *speaker adaptation* dengan menggunakan *speech manipulation tools* yaitu *STRAIGHT* (*Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*) dan DNN (deep neural network) yang disediakan oleh HTS *demo speaker adaptation* terbaru dengan keseluruhan sistem yang baru.

Halaman ini sengaja dikosongkan

DAFTAR PUSTAKA

Black, A. W., & Campbell, N. (1996). OPTIMISING SELECTION OF UNITS FROM SPEECH DATABASES FOR CONCATENATIVE SYNTHESIS.

Cahyaningtyas. (2018). *Speaker Adaptation Pada Sistem Sintesis Ucapan Bahasa Indonesia Berbasis Hidden Markov Model*. Surabaya: Institut Teknologi Sepuluh Nopember.

Cahyaningtyas, E. (2015). *Speech Synthesis Bahasa Indonesia Berbasis Hidden Markov Model (HMM) Pada Intonasi Kalimat Berita dan Kalimat Tanya.* Surabaya: Institut Teknologi Sepuluh Nopember.

Gao, Y. (2001). Electronic Braking System of EV and HEV--Integration of Regenerative Braking, Automatic Braking Force Control and ABS. *42 Volt Technology and Advanced Vehicle Electrical Systems*.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for. *Computer Speech and Language*, 171–185.

Lestari, D. M. (2018). *Average Voice Model (AVM) Sintesa Alamiah Bahasa Indonesia Berbasis Hidden Markov Models (HMM).* Surabaya: Institut Teknologi Sepuluh Nopember.

Lida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion . *Speech Communication 40*, 161-187.

Lubbers, K. (2014). Design and Analysis of a Model Based Low Level Slip Controller Based on a Hybrid Braking System. *Science in Systems and Control Delft University*.

Moulines, E., & Charpentier, F. (1990). PITCH-SYNCHRONOUS WAVEFORM PROCESSING TECHNIQUES FOR TEXT-TO-SPEECH SYNTHESIS USING DIPHONES . *Speech Communication 9*, 453-467.

Murali, T. (2017). Four Quadrant Operation and Control of Three Phase BLDC Motor. *International Conference on Circuits Power and Computing Technology*.

Povey, D., & Yao, K. (2012). A Basis Representation of Constrained MLLR. *Elsevier, Computer Speech and Language*, 35-51.

Singh, C. P. (2012). State-space Based Simulink Modeling of BLDC Motor and its Speed Control Using Fuzzy PID Controller. *International Journal of Advances in Engineering Science and Technology*, *2*, 359-369.

Tashakori, A. (2011). Modeling of BLDC Motor with Ideal Back-EMF for Automotive Applications. *World Congress on Engineering.* London.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE*. Japan.

Tur, O. (2007). An Introduction to Regenerative Braking of Electric Vehicles as Anti-Lock Braking System. *Proceedings of 2007 IEEE Intelligent Vehicles Symposium*, (pp. 13-15). Istanbul.

Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis.*

Yamagishi, J., Masuko, T., Tokuda, K., & Kobayashi, T. (2003). A TRAINING METHOD FOR AVERAGE VOICE MODEL BASED ON SHARED DECISION. *ICASSP*, 716-719.

Zen, Heiga, Tokuda, K., & Alan, B. W. (2009). Statistical parametric speech synthesis. *Speech Communication 51.11*, pp. 1039-1064.

LAMPIRAN

A. Langkah-langkah Menjalankan *Software HMM-based Speech Synthesis System Speaker Adaptation*

1. Nyalakan komputer dengan sistem operasi Ubuntu
2. Pastikan pada komputer sudah menginstall semua tools yang dibutuhkan
3. Buka terminal pada lokasi/folder yang ingin di *run*
4. Panggil *Envars Tools*:
   export TOOLS_DIR=~/.hts_sptk
   export PATH=$TOOLS_DIR/bin:$PATH
   export PATH=$TOOLS_DIR/festival/bin:$PATH
   export PATH=$TOOLS_DIR/speech_tools/bin:$PATH
   export FESTVOXDIR=$TOOLS_DIR/festvox
   export FESTDIR=$TOOLS_DIR/festival
   export ESTDIR=$TOOLS_DIR/speech_tools
   export PATH=$FESTDIR/examples:$PATH

5. Konfigurasi alamat *tools* untuk *script* demo:
   chmod a+x configure
   ./configure \
   --with-fest-search-path=$TOOLS_DIR/festival/examples \
   --with-sptk-search-path=$TOOLS_DIR/bin \
   --with-hts-search-path=$TOOLS_DIR/bin \
   --with-hts-engine-search-path=$TOOLS_DIR/bin \
   DATASET=vibid ADAPTSPKR=mmht ADAPTHEAD=15

6. Sesuaikan data/makefile dan scripts/config.pm dengan basis data sintesis yang diinginkan
7. Persiapan data dan skrip
   make all 2>&1 | tee log_prepare_$NMFILE.txt

8. Sintesis kalimat:
   scriptpath=$(pwd) perl scripts/Training.pl $scriptpath/scripts/Config.pm 2>&1 | tee log_synthesis_$NMFILE.txt

B. Kode Matlab Untuk Mencari Frekuensi Dasar

```
clear; clc;
[X,Fs]=audioread('C:\Users\Farahiyah\Documents\SIAP_SIDANG\HASIL SUARA\ind_2\SI\0\vibid_mmht_vibid_mmht_0100.wav');
f0=exstraightsource(X,Fs);
d=diff(f0,2);
time = (1:length(X))/Fs*1000;

figure;
subplot(1,1,1);
plot(f0);
title('F0');
xlabel('Time (ms)'); ylabel('Frequency (Hz)');
```

Halaman ini sengaja dikosongkan

BIODATA PENULIS

Penulis dilahirkan di Jakarta, pada tanggal 10 September 1998 dan merupakan anak kedua dari dua bersaudara. Penulis telah menyelesaikan pendidikan formal dari SDS Sumbangsih Grogol, SPMN 45 Jakarta, SMAN 78 Jakarta, dan terakhir di Teknik Fisika FTIRS-ITS Program Studi S1. Selama masa kuliah , penulis mengambil bidang minat Vibrasi dan Akustik dan aktif menjadi asisten Laboratorium Vibrasi dan Akustik. Penulis pernah melakukan magang di PT. Telkom Indonesia sebagai wireless network intern. Organisasi yang pernah diikuti penulis adalah sebagai project manager di AIESEC Surabaya. Organisasi minat dan Bakat yang pernah diikuti oleh penulis adalah sebagai staf UKM Perisai Diri. Hobi penulis adalah olahraga, membaca, dan menulis. Penulis dapat dihubungi via email farahiyahaisahsidik@gmail.com.