Aalto University

Licentiate Thesis

A wave equation model for vowels: Measurements for validation

Pertti Palo

Supervisor: Professor Timo Eirola
Advisor: Dr. Tech. Jarmo Malinen

Espoo
2.6. 2011

AALTO UNIVERSITY

ABSTRACT OF THE LICENTIATE THESIS

Author:	Pertti Palo
Title of the thesis:	A wave equation model for vowels: Measurements for validation
Date:	2.6. 2011 **Number of pages:** 88
School:	School of Science
Professorship:	Mat-1
Supervisor:	Professor Timo Eirola
Instructor:	Dr. Tech. Jarmo Malinen

This thesis presents a procedure for simultaneous recording of anatomic and acoustic data on speech production. The thesis also presents a mathematical model of vowel production and its pilot implementation as frequency domain numerical simulator. The model is based on the wave equation.

The anatomic data is recorded with MRI and the acoustic data with a purpose built, MRI compatible sound recording setup. Both modalities of the data will be used in building the next generation of the simulator, estimating its parameters, and validating it. Compared with previous contributions in this area, the acoustic and anatomic data are more closely matched and, hence, the procedure more informative. The main scientific contribution is the description of the sound recording arrangement as well as the guidelines and considerations detailed for simultaneous data acquisition in MRI.

Keywords: Articulatory speech synthesis, FEM, finite element method, MRI, speech recording in noise, speech recording in MRI, vowel synthesis, wave equation

AALTO YLIOPISTO

LISENSIAATTITYÖN TIIVISTELMÄ

Tekijä:	Pertti Palo
Työn nimi:	Aaltoyhtälömalli vokaaleille: Mittauksia validointia varten
Päivämäärä:	2.6. 2011 **Sivuja:** 88
Korkekoulu:	Perustieteiden korkeakoulu
Professuuri:	Mat-1
Työn valvoja:	Professori Timo Eirola
Työn ohjaaja:	Tekn. Tri. Jarmo Malinen

Tämä työ esittelee järjestelmän, jolla voidaan kerätä yhtäaikaista akustista ja anatomista dataa puheen tuotosta. Työ esittelee myös vokaalituoton matemaattisen mallin ja tämän mallin pilottitoteutuksen numeerisena resonanssisimulaattorina. Malli perustuu aaltoyhtälöön.

Anatominen data kerätään MRI:llä ja akustinen data erikseen tätä tarkoitusta varten rakennetulla äänitysjärjestelmällä. Dataa käytetään seuraavan simulaattorisukupolven rakentamisessa, sen parametrien estimoimisessa ja validoinnissa. Verrattuna aiempiin vastaaviin järjestelmiin, akustinen ja anatominen data vastaavat tarkemmin toisaan ja siksi data on informatiivisempaa. Tärkein tieteellinen anti on äänitysjärjestelmän kuvaus sekä suositukset ja huomiot yhtäaikaisesta datankeruusta MRI:n aikana.

Avainsanat: Aaltoyhtälö, artikulatorinen puhesynteesi, elementtimenetelmä, FEM, MRI, puheen äänitys melussa, puheen äänitys MRI:ssä, vokaalisynteesi

To absent and unmet friends

Acknowledgements

This Licentiate thesis has been done for the Institute of Mathematics of Aalto University. It has been partially funded by the Instrumentarium Science Foundation, the Cultural Foundation of Finland, the Institute of Mathematics, and in its final stages by my parents Helena and Veikko Palo.

I want to thank my supervisor professor Timo Eirola, who has helped me along with my studies, and my instructor Jarmo Malinen, who among other things has provided a wealth of comments, taught me a lot about measurement projects and the construction and calibration of measurement equipment.

A special thank you is also due to Stina Ojala and Micheal O'Dell for reading the thesis and providing comments on it.

I would also like to thank my friends including Daniel Aalto, Riikka Kangaslampi, Stina Ojala, and Anu Leponiemi as well as several others. You all have listened to me ranting about this thesis, angsting about it and so on... And then done it again the next evening. Thank you.

Finally, my gratitude goes to my family. The upbringing I got, the ongoing support from my parents and having a kid sister – and a brother-in-law and a nephew – are very important to me.

Laajalahti, June 2, 2011 Pertti Palo

Contents

Acknowledgements iv

Contents v

List of Publications vi

List of Figures vii

List of Tables xi

Symbols and Abbreviations xii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Current context . . . 2
- 1.3 Model construction . . . 3

2 Mathematical model of vowel production 6
- 2.1 Acoustic model . . . 7
- 2.2 Notes on deriving the acoustic model . . . 8
- 2.3 Eigenvalue problem and variational formulation . . . 9

3 Computational model of vowel production 11
- 3.1 Computational resonance model . . . 11
  - 3.1.1 Simulation practicalities . . . 12
- 3.2 Simulated formants . . . 13
- 3.3 Accuracy of the simulation . . . 14

4 Experimental design 16
- 4.1 Requirements . . . 16
- 4.2 Experimental setting . . . 18
- 4.3 Phonetic materials and subject . . . 19

4.4 MRI sequence 20
4.5 Sound measurements on long phonations 21

5 Sound recording arrangement 23
5.1 Principle of noise cancellation 24
5.2 Physical recording setup 26
5.2.1 Sound collector 27
5.2.2 Acoustic wave guides 29
5.2.3 Shielded microphone array 30
5.2.4 De-noising amplifier 31
5.3 Response measurements 32
5.4 Computer equipment and digital signal processing 36

6 Measurement results 37
6.1 Sound data from MRI 37
6.1.1 f_0 and formant extraction 37
6.1.2 Spectral analysis 39
6.1.3 Acoustic noise data 40
6.2 MRI data 44
6.2.1 Data for developing a set of disqualification criteria . 45
6.3 Stability data on long vowel productions 47

7 Discussion 51
7.1 Observations on the results 51
7.1.1 Modeling 51
7.1.2 Recording setup 52
7.1.3 Sound data from MRI 52
7.1.4 Anatomical data 53
7.1.5 Stability of long vowel productions 54
7.2 Recommendations and future directions 54
7.2.1 Model validation 54
7.2.2 Data acquisition 55
7.2.3 Other observations 56
7.3 Conclusion 57

Bibliography and References 59

A Sound data 63

B Deviation data on long vowel productions 82

List of Publications

Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (2006). Formants and vowel sounds by finite element method. In *The Phonetics Symposium 2006*, pages 24 – 33, Helsinki, Finland.

Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (2007). Vowel formants from the wave equation. *Journal of the Acoustical Society of America Express Letters*, 122(1):EL1–EL7.

Lukkari, T., Malinen, J., and Palo, P. (2007). Recording speech during magnetic resonance imaging. In *MAVEBA 2007*, pages 163 – 166, Florence, Italy.

Lukkari, T., Malinen, J., and Palo, P. (2008). Puheen äänittäminen magneettiresonanssikuvauksen aikana. In *Fonetiikan päivät 2008*, pages 57 – 64, Tampere, Finland.

Malinen, J. and Palo, P. (2009). Recording speech during MRI: Part II. In *MAVEBA 2009*, pages 211–214, Florence, Italy.

Aalto, D., Malinen, J., Palo, P., Aaltonen, O., Vainio, M., Happonen, R.-P., Parkkola, R., and Saunavaara, J. (2011a). Recording speech sound and articulation in MRI. In *Biodevices 2011*, pages 168 – 173, Rome, Italy.

Aalto, D., Malinen, J., Palo, P., Saunavaara, J., and Vainio, M. (2011b). Estimates for the measurement and articulatory error in MRI data from sustained phonation. To appear in *Proceedings of ICPhS 2011*, Hong Kong, China.

List of Figures

1.1 Constructing a scientific model 3
1.2 Constructing a mathematical model 4
1.3 Constructing a physical model 5

3.1 Cross-sectional representation of the vocal tract 12
3.2 F1-F4 pressure distributions 14
3.3 Isobars of F1-F4 14
3.4 Computed and measured vowels in the (F2,F1)-plane. 15

4.1 Test subject and equipment under preparation 18
4.2 Annotated spectrogram of a sound recording 20
4.3 Measuring long phonations in the anechoic chamber 22

5.1 Sound recording schematic 24
5.2 Sound collector measurement 26
5.3 (a) From left to right: the face model, a reference microphone and the sound collector (b) Prototypes for a paraboloid reflector (c) One of the paraboloid reflectors suspended from a prototype support structure. 27
5.4 Sound collector from the noise channel side 28
5.5 Sound collector from the speech channel side 28
5.6 The acoustic wave guides, stative and the sound collector . . 29
5.7 The microphone array inside the Faraday cage. 30
5.8 Microphone array 31
5.9 De-noising amplifier 32
5.10 A mock up of a noise channel frequency response measurement. 33
5.11 Frequency response of the speech channel with and without the face model 33
5.12 The effect of a microphone-sized object on the frequency response of the speech channel. 34

5.13 Frequency responses of the speech and noise channel 35

6.1 Formant shifts of the MRI data set 38
6.2 The speech channel's frequency response as used in spectral compensation of the smoothened spectra 40
6.3 [ɑ] target f_0 =110 Hz. MRI sequence VIBE 1.8, duration 7.6 s. 41
6.4 [æ-ɑ] target f_0 =110 Hz. MRI sequence VIBE 1.8, duration 7.6 s. 42
6.5 Acoustic noise charasteristics of a 8 s dynamic MRI sequence 43
6.6 Acoustic noise charasteristics of the VIBE 1.8 MRI sequence 43
6.7 Sagittal section of [æ] 44
6.8 Sagittal section of [y] 44
6.9 Sagittal section of [æ-ɑ] glide 45
6.10 Sagittal sections of [ɑ] 46
6.11 Sagittal sections of a [æ-ɑ] glide 46
6.12 Sagittal sections of a [æ] 47
6.13 Optimal sampling time distribution for the whole data. . . . 48

A.1 [ɑ] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6s. . . . 64
A.2 [e] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 65
A.3 [i] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . . 66
A.4 [o] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 67
A.5 [u] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 68
A.6 [y] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 69
A.7 [æ] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 70
A.8 [ø] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 71
A.9 [ɑ] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 72
A.10 [e] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 73
A.11 [i] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . . 74
A.12 [o] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 75
A.13 [u] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 76
A.14 [y] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 77
A.15 [æ] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 78
A.16 [ø] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s. . 79
A.17 [æ-ɑ] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s. . 80
A.18 [æ-ɑ] target f_0 =110 Hz, sequence: dynamic, duration 8 s. . . 81

B.1 Short and long phonations with both target f_0s.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4. 83

B.2 Duration under 16 s.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4. 84
B.3 Duration over 16 s.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4. 85
B.4 Target f_0 = 110 Hz.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4. 86
B.5 Target f_0 = 137.5 Hz.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4. 87
B.6 Optimal sampling time distribution for the whole data. . . . 88
B.7 Optimal sampling time distributions for different fractions of the data. 88

List of Tables

3.1 Formants for [øː] 13

4.1 The MRI sequencies and their parameters. 21

6.1 Optimal beginning times (in seconds) for a 7.6 s sample. . . 48
6.2 f_0s, formants F1-F4, formant distances and sound pressure level differences for vowel productions with target $f_0 = 110$ Hz. Glides with [1] Vibe 1.8 and [2] dynamic imaging sequences. Aliased F2 values are shown in [square brackets . . 49
6.3 f_0s, formants F1-F4, formant distances and sound pressure level differences for vowel productions with target $f_0 = 137.5$ Hz. Aliased F2 values are shown in [square brackets . 50

Symbols and Abbreviations

$\frac{\partial}{\partial x}$ partial derivative with respect to x

$\frac{\partial^2}{\partial x^2}$ second partial derivative with respect to x

$\frac{\partial}{\partial \nu}$ partial derivative with respect to ν (see below)

$A(x)$ area function i.e. cross sectional area of the VT in relation to distance from glottis

ν exterior unit normal of a domain

ρ_{Tot} total density of the medium that is $\rho_{Tot} = \rho_0 + \rho$

ρ_0 constant component of the density of the medium

ρ perturbation of the density of the medium

Δ the Laplace operator: $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$

Φ velocity potential

Φ_t first time derivative of velocity potential

Φ_{tt} second time derivative of velocity potential

P total pressure that is $P = p_0 + p$

Ω a domain in $\mathbb{R}^3$

$\partial\Omega$ boundary of Ω

c speed of sound

p_0 static pressure (atmospheric pressure)

p perturbation pressure (sound pressure)

q' mass source

u pressure input signal at the glottis

U volume velocity

$\overline{u}$ particle velocity

$[\mathrm{a}]$ phone a

1D One Dimensional

2D Two Dimensional

3D Three Dimensional

C Consonant (e.g. CV sequence = consonant-vowel sequence)

CMRR Common Mode Rejection Ratio

DSP Digital Signal Processing

F0 Fundamental Frequency

F1, F2, F3, ... Formant Frequencies

FEM Finite Element Model

fMRI functional Magnetic Resonance Imaging. **Not** used in this work.

FOV Field Of View

FFT Fast Fourier Transform

IPA International Phonetic Association

LPC Linear Predictive Coding

MR Magnetic Resonance (as in MR image)

MRI Magnetic Resonance Imaging

SNR Signal to Noice Ratio

V Vowel (e.g. VCV sequence = vowel-consonant-vowel sequence)

VF Vocal Folds

VT Vocal Tract

Chapter 1

Introduction

This work presents a procedure for recording anatomic and acoustic data on speech production. The data is used in building a speech production simulator, estimating its parameters, and validating it. Both modalities (anatomic and acoustic) are acquired simultaneously. Recording the sound and geometry as a matching pair makes the data more reliable.Compared with previous contributions in this area, the acoustic and anatomic data are more closely matched. The main scientific contribution is the description of the sound recording arrangement in Chapter 5 as well as the guidelines and considerations detailed in Chapter 7.

1.1 Background

Our main goal is to understand human speech production. As part of the process of testing our understanding and exploring the relevant phenomena we are developing mathematical and numerical models of speech production. The models are based on the acoustic theory of speech production or source-filter theory as layed out by Helmholtz (1863), Chiba and Kajiyama (1941), and Fant (1960). Consequently, the vocal tract modeling does not incorporate any nonlinearities. As such, the models will be applicable to vowel and nasal production, while most of the consonant classes will lie outside their scope for now.

Even such a relatively simple model can be used for many purposes – if it is implemented and validated carefully. In phonetics it can be used to explore questions of vowel prosody, timing in diphthong production, and speaker characteristics. More practical applications can be found in speech technology where inverse filtering and speech recognition, see e.g. Blackburn (1996), as well as speaker recognition can benefit from employing the

model.

In medicine, such a model can be used in, for example, planning oral and maxillofacial surgery and in studying the effects of abnormal or altered anatomy (Švancara and Horáček 2006). Mathematical modeling of surgical operations is quickly becoming reality, see e.g. Deufhard et al. (2006). While a 'surgical speech simulator' is still a long way off, the data acquisition procedure presented here can already be used in assessing the changes in a patient's speech by comparing pre- and post-operation data.

Let us consider the question:"To what extent can the acoustic theory of speech production be considered accurate?" There certainly is a limit to its applicability. Consider, for example, a Finnish [r] and sounds produced with the aryepiglottal vocal folds, for the latter, see e.g. Moisik (2008). Both require the acoustical theory to be amended by flow dynamics. However, it seems that a wide range of speech sounds should be accurately modelled by linear acoustics (given a separate vocal source): vowels, nasals, stable liquids such as [l] and even some approximants with short duration such as the Finnish [v]. And, of course, the transitions between any of these sounds or more accurately fluent speech made up of these sounds. Still, there are cases when the source-filter theory is insufficient even for vowels: Within transitions between sound registers (e.g. from chest register to falsetto) the source and filter are not always separable (Titze 2008).

1.2 Current context

There are no accurate methods to measure speech production as a whole. Accurate here means repeatable and exact. Even if we look at the physics of speech production only and leave out cognitive, neurological and most of the physiological factors, this statement is still true. However, there are several accurate ways to measure parts of speech production. To give two examples: recording the output sound signal in an anechoic room, and performing inverse filtering on such a signal to acquire the glottal source signal.

This work is a step towards the goal of measuring speech production in a repeatable and exact way. As it is, we do not yet have the final solution, but we have a very good starting point and have made considerable progress towards our goal.

Model construction consist necessarily of a refinement process. At first, it is difficult to know what exactly should be measuredn and modeled and how to do it. When starting the process, we have only a fragmented and partial understanding of the phenomena of interest. Our theories and

concepts concerning the topic are fuzzy and incomplete. By conducting experiments and constructing models based on them, we understand the situation in more detail and our concepts become more exact. This, in turn, clarifies the question of what to measure and how. Each repetition of the process makes our measurements more repeatable and thus it can be said that our understanding has grown: our theory has become more exact.

1.3 Model construction

Roughly speaking, the method of constructing a scientific model can be said to consist of three interdependent elements. These are *1. measurements*, which guide the construction and validation of *2. models* which guide the formation and understanding of *3. concepts and theories*. This, in turn, guides the planning and implementation of measurements and so on... This can be presented as the flow chart in Figure 1.1.

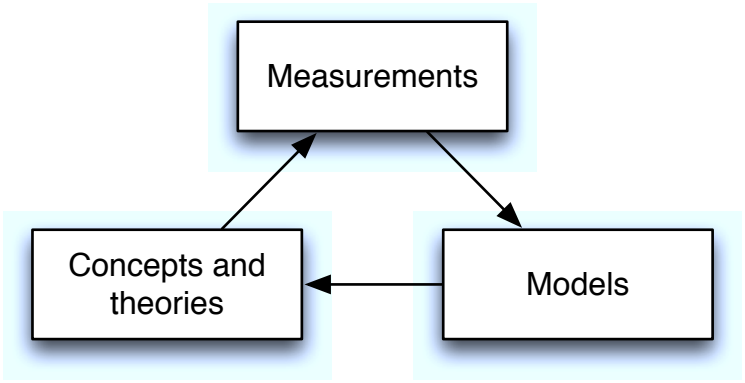


Figure 1.1: The process of constructing a scientific model

As said, it is a rough representation of the process. A more refined flow chart is shown in Figure 1.2. It presents the construction process of a mathematical model – such as ours – as divided into five stages. Here the process is refined to start with the stages of *1. Problem definition* and *2. System analysis*. In practice this means that we answer the questions "What is the problem?", "What is the system associated with the problem?" and "What is relevant?"

Next, we have the stages of *3. Modeling* and *4b. Simulation*. These stages give a mathematical formulation for our model and implement it as numerical computer code. The code needs data from the phenomenon

as its basis before a simulation run. This is provided by stage *4a. Data acquisition*, which may consist of more than one stage. Data acquisition is also closely linked to the final stage *5. Validation*. Together with it they form the measurements block of the first flow chart.

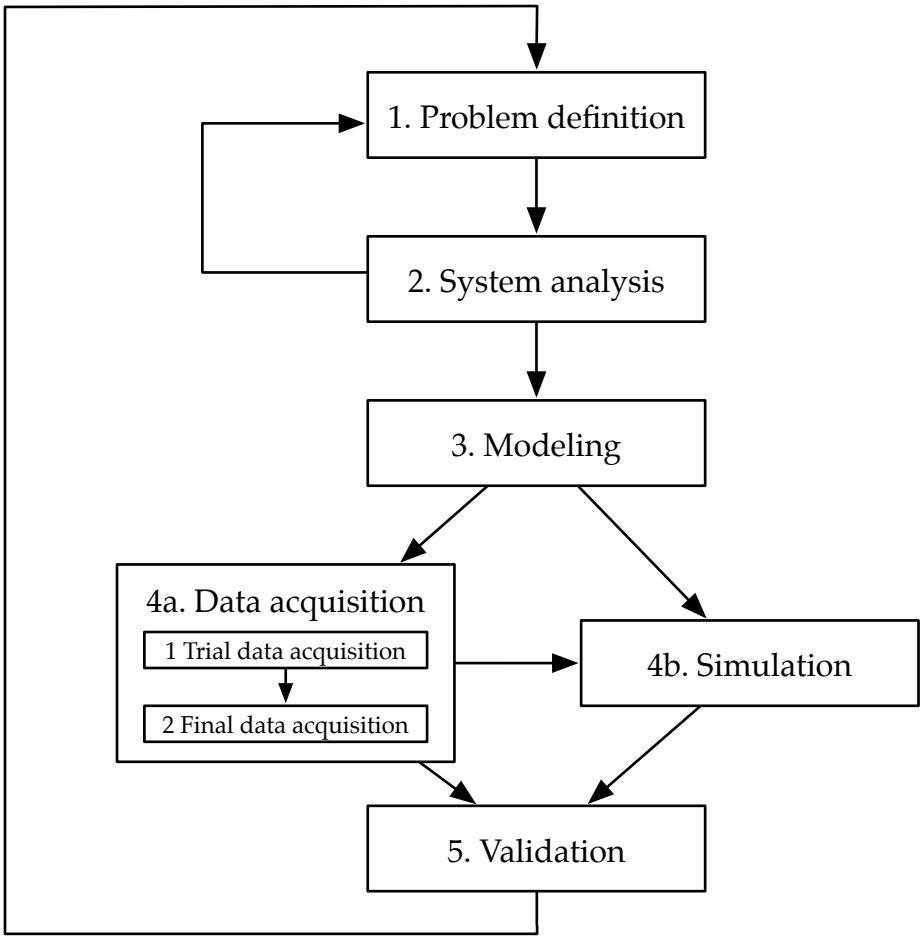


Figure 1.2: The process of constructing a mathematical model

From the point of view of this work, stage 4a in Figure 1.2 consists of measuring and validating the behaviour of the measurement equipment itself. The process involves construction of a physical model of the speaker and this part is expanded in some detail in Figure 1.3. The stages are quite similar to those of Figure 1.2. However, some differences should be pointed out. Stage *3. Modeling* consists of the construction of an actual physical object: The face model. Stage *4a Simulation* means frequency response measurements of the sound recording system in an anechoic chamber and Stage *4b Data acquisition* is the pilot data acquisition in the actual

MRI environment.

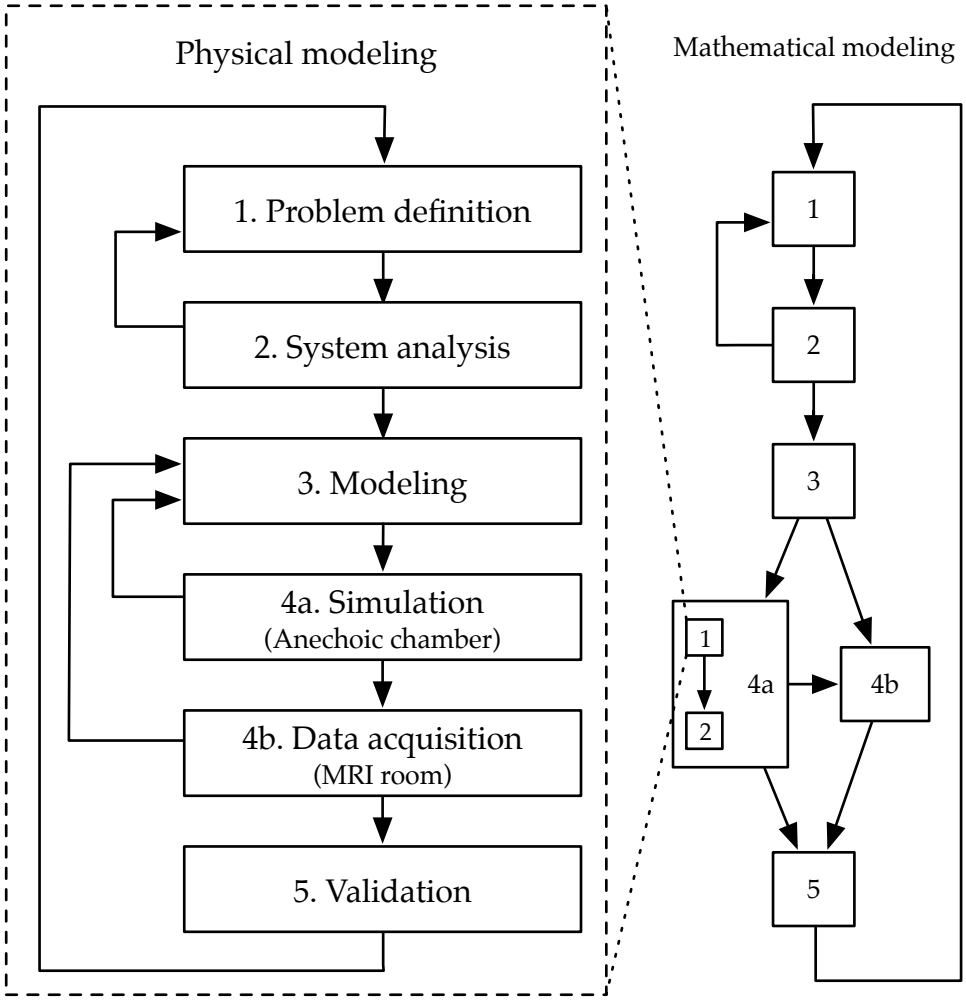


Figure 1.3: The process of constructing physical model

Even the chart in Figure 1.3 is lacking in detail. Also the human subject needs to be measured to arrive at the best possible data acquisition plan for the mathematical model construction which, after all, is the actual thing we are interested in. As can be seen, constructing an accurate flow chart of the modeling process is a rather complex and difficult task which can be iterated ad infinitum. Rather than presenting yet one more flow chart let us turn our attention to the actual modeling work. We will start with problem definition, system analysis and the formulation of the mathematical model in the next Chapter.

Chapter 2

Mathematical model of vowel production

Our mathematical and computational models are based on the acoustic theory of speech production (Fant 1960). This theory separates the glottal sound source from the filter, formed by the vocal tract, and treats them as independent subsystems. In this paradigm, a vowel is produced when a relatively open vocal tract filters the sound produced by the glottis[1].

According to the source-filter paradigm we have separated the glottal source and vocal tract (VT) filter into different modules (or submodels). However, the modeling focus of this thesis is with the time independent resonance modeling, and therefore the glottal model can be left mostly out of the discussion. The glottal model is reported by Aalto (2009) and Aalto et al. (2009).

Our mathematical model, see Equation (2.2), captures many relevant phenomena of wave propagation in three-dimensional geometry (e.g., to detect cross modes). However, its underlying assumptions exclude non-linear phenomena such as turbulence and shock formation, or losses due to viscosity, heat conduction, or boundary dissipation. While the absence of non-linear phenomena in the vocal tract is in accordance with the source-filter paradigm, the exclusion of losses is a further simplification.

Our mathematical and computational models of vowel production and the modeling results are also reported in Hannukainen et al. (2006; 2007). This and the next chapter follow the discussion presented in that article.

[1] or by noise generated at the glottis if the subject is whispering

2.1 Acoustic model

The geometry of our acoustic model consists of the interior of the VT $\Omega \subset \mathbb{R}^3$ and its boundary $\partial\Omega$. The boundary is made up by three parts with different types of behaviour: $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$. The surface Γ_1 is the mouth opening, Γ_2 stands for the walls of the VT, and Γ_3 is the glottis end of the VT.

Specifically, the glottis end Γ_3 is a boundary plane between the vocal tract (filter) model and the glottal (source) model. It is not motivated by anatomy, but rather by physics. In other words, there is no anatomical division at the point we have chosen, but rather a transition from the area close to glottis in terms of physical phenomena to the area relatively far from it.

In contrast, the wall Γ_2 of the vocal tract is a clear anatomical boundary. The behaviour we have chosen for it presents it as an infinitely hard barrier which reflects all incoming sound energy back into the VT.

The mouth opening Γ_1 is more of a mixed case. While there is a clear conceptual boundary where a finite tube (the VT) opens to a practically infinite space (the space around the speaker), its precise placement is not so clear. In fact, it should probably be made frequency dependent. As a first approximation, we have chosen to model it simply at one, heuristically chosen surface where we assume that the VT opens abruptly to the surrounding space.

The interior space Ω of the VT is modeled by the wave equation. To start deriving the wave equation for sound pressure, we assume that the total pressure $p = p(\mathbf{r}, t)$ and the density $\rho = \rho(\mathbf{r}, t)$ can be expressed as a sum of a static part and a time-dependent perturbation:

$$p(\mathbf{r}, t) = p_0 + p'(\mathbf{r}, t) \quad \text{and} \quad \rho(\mathbf{r}, t) = \rho_0 + \rho'(\mathbf{r}, t), \tag{2.1}$$

respectively, where p_0 and ρ_0 are independent of time t and the space variable $\mathbf{r}$. For linearisation of the model, we assume that $p' = p'(\mathbf{r}, t) \ll p_0$ and $\rho' = \rho'(\mathbf{r}, t) \ll \rho_0$ are small perturbations at point $\mathbf{r} = (x, y, z) \in \Omega$ at time t.

The velocity field $\mathbf{v} = \mathbf{v}(\mathbf{r}, t)$ of the air movement is described by p and ρ. A velocity potential $\Phi = \Phi(\mathbf{r}, t)$ is any function that satisfies $\mathbf{v} = -\nabla\Phi$. With this notation, our acoustic model is given by

$$\begin{cases} \Phi_{tt} = c^2\Delta\Phi & \text{for } (\mathbf{r}, t) \in \Omega \times \mathbb{R}, \\ \Phi = 0 & \text{for } (\mathbf{r}, t) \in \Gamma_1 \times \mathbb{R}, \\ \frac{\partial\Phi}{\partial\nu} = 0 & \text{for } (\mathbf{r}, t) \in \Gamma_2 \times \mathbb{R}, \text{ and} \\ \Phi_t + c\frac{\partial\Phi}{\partial\nu} = 2\sqrt{\frac{c}{\rho_0}}u & \text{for } (\mathbf{r}, t) \in \Gamma_3 \times \mathbb{R}, \end{cases} \tag{2.2}$$

where $u = u(\mathbf{r}, t)$ is the incoming power per unit area at the glottis end, c is the sound velocity in the VT, ν is the exterior unit normal on $\partial\Omega$, and $\frac{\partial\Phi}{\partial\nu} = \nu \cdot \nabla\Phi$. The task is to compute the velocity potential $\Phi(\mathbf{r}, t)$ for a given glottal input $u(\mathbf{r}, t)$.

2.2 Notes on deriving the acoustic model

To derive Equations (2.2) from "first principles", one needs to assume that an isentropic thermodynamic equation of state for pressure $p = p(s, \rho)$ holds where s, ρ are the entropy and density, respectively. Then we define the speed of sound c by linearising the equation of state

$$p' = p(s, \rho_0 + \rho') - p(s, \rho_0) \approx c^2\rho'$$

where $p_0 = p(s, \rho_0)$ and $c^2 = \frac{\partial p}{\partial \rho}(s, \rho_0)$. In this approximation, the entropy s is kept constant since the associated thermodynamic process is assumed to be (locally) reversible. In the case of monatomic ideal gas, we have $p/\rho^\gamma = p_0/\rho_0^\gamma$ and $c^2 = \gamma p_0/\rho_0$ where $\gamma = 5/3$ is the adiabatic constant.

Now the wave equation $\Phi_{tt} = c^2\Delta\Phi$ can be derived by a linearisation argument involving the continuity equation, Euler equation and linearised equation of state $p' = c^2\rho'$. Having computed Φ, we obtain the perturbation pressure from

$$p' = \rho_0\Phi_t. \tag{2.3}$$

All this can be found, e.g., in Chapter 9 of Fetter and Walecka (1980)

We also need to take into account the walls and both ends of the VT. The last three lines in Equations (2.2) specify the required boundary conditions. We regard the mouth as an open end of an acoustic tube, and this is modelled by the Dirichlet condition $\Phi(\mathbf{r}, t) = 0$.

On the walls of the VT, we use the same Neumann condition $\frac{\partial\Phi}{\partial\nu}(\mathbf{r}, t) = 0$ as one would use at the closed end of a resonating tube. These two boundary conditions are discussed by Fetter and Walecka (1980: pp. 306–307).

At the glottis, we use a scattering boundary condition that specifies the ingoing sound *energy* wave. For motivation, define the ingoing wave $u(\mathbf{r}, t)$ and the outgoing wave $y(\mathbf{r}, t)$ for $\mathbf{r} \in \Gamma_3$ by

$$u = \sqrt{\frac{\rho_0}{4c}}\left(c\frac{\partial\Phi}{\partial\nu} + \Phi_t\right) \quad \text{and} \quad y = \sqrt{\frac{\rho_0}{4c}}\left(c\frac{\partial\Phi}{\partial\nu} - \Phi_t\right). \tag{2.4}$$

The first of these equations coincides with the third boundary condition in (2.2). The net power absorbed by the interior domain Ω through the

control/observation boundary at time t satisfies

$$\int_{\Gamma_3} |u(\mathbf{r},t)|^2 \, d\omega(\mathbf{r}) - \int_{\Gamma_3} |y(\mathbf{r},t)|^2 \, d\omega(\mathbf{r}) = \int_{\Gamma_3} (-\nu(\mathbf{r})) \cdot \mathbf{j_e}(\mathbf{r},\mathbf{t}) \, d\omega(\mathbf{r})$$

where $\mathbf{j_e} = -\rho_0 \Phi_t \nabla\Phi = p'\mathbf{v}$ is the energy-flux vector as introduced in Fetter and Walecka (1980: pp. 307). Please note that we have a minus sign in front of the exterior unit vector ν because we regard the energy coming into Ω as positive (and the outgoing as negative).

2.3 Eigenvalue problem and variational formulation

Simulating vowels based on Equations (2.2) would require specifying a glottal input signal *field* at the glottal boundary Γ_3. While several glottal source models exist, adapting one to provide an input for our vocal tract model is a non-trivial task. This is caused by the fact, that our model requires as its input the velocity potential over the glottal surface Γ_3 while most glottal models provide a scalar volume velocity as their output, see e.g. Fant et al. (1986) and Ishizaka and Flanagan (1972). A forward coupling (glottal model to VT model) is demonstrated by Alku et al. (2006). However, a feedback coupling (VT model to glottal model) is harder to implement between such models.[2]

However, there is another way to look at the problem. The approach actually rises from the fact that in a phonetics research context, vowels are usually defined by their sonorant quality and the peaks within their spectrum. A speech sound is sonorant if is produced while the vocal folds vibrate. The spectral peaks of such a sound are called *formants*. In the source-filter paradigm, the formants are thought to correspond to the resonances of the VT.

So, instead of solving the time-dependent Equations (2.2), we can solve an easier, related, problem. We determine the resonance frequencies corresponding to a particular vowel articulation position. By e.g. Theorem 2.3 of Malinen and Staffans (2006), the resonances of Equations (2.2) can be solved by finding the discrete, complex frequencies λ and the cor-

[2]For an example of a 1D feedback coupled model, see Aalto (2009). We will use this model for comparisons when developing a 3D feedback coupled model.

responding nonzero eigenfunctions $\Phi_\lambda(\mathbf{r})$ such that the equations

$$\begin{cases} \lambda^2\Phi_\lambda = c^2\Delta\Phi_\lambda & \text{on } \Omega, \\ \Phi_\lambda = 0 & \text{on } \Gamma_1, \\ \frac{\partial\Phi_\lambda}{\partial\nu} = 0 & \text{on } \Gamma_2, \text{ and} \\ \lambda\Phi_\lambda + c\frac{\partial\Phi_\lambda}{\partial\nu} = 0 & \text{on } \Gamma_3 \end{cases} \tag{2.5}$$

are satisfied. The time harmonic extension $\Phi(\mathbf{r}, t) = \Phi_\lambda(\mathbf{r})e^{\lambda t}$ of Φ_λ satisfies Equations (2.2). Using Equation (2.3), the corresponding perturbation pressure distribution is given by $p'(\mathbf{r}, t) = p_\lambda(\mathbf{r})e^{\lambda t}$ where $p_\lambda(\mathbf{r}) := \rho_0\lambda\Phi_\lambda(\mathbf{r})$. Thus Equations (2.5) are satisfied with p_λ in place of Φ_λ, as well.

Finally, to provide a basis for simulation by the Finite Element Method (FEM), we compute the variational formulation of Equations (2.5) (with p_λ in place of Φ_λ):

$$\lambda^2 \int_\Omega p_\lambda\phi \,\mathrm{d}\Omega + \lambda c \int_{\Gamma_3} p_\lambda\phi \,\mathrm{d}\omega + c^2 \int_\Omega \nabla p_\lambda \cdot \nabla\phi \,\mathrm{d}\Omega = 0, \tag{2.6}$$

where ϕ is an arbitrary test function in the Sobolev space $H^1_{\Gamma_1}(\Omega) = \{f \in H^1(\Omega) : f(\mathbf{r}) = 0 \text{ for } \mathbf{r} \in \Gamma_1\}$. The formulation incorporates the boundary conditions on Γ_1 (explicitly) and Γ_2 (implicitly). The boundary condition on Γ_3 (glottis end) is presented by the second term of the variational formulation.

Chapter 3

Computational model of vowel production

This Chapter introduces the results of a limited practical test of the synthesis concept. Starting in 2006, we used the data Associate Professor Olov Engwall (KTH) most helpfully gave us (Engwall and Badin 1999). Thus, we could explore our concept for the simulator without any data aqcuisition of our own.

3.1 Computational resonance model

The Finite Element Method (FEM) can be used to approximately solve Equation (2.6); see, e.g., Johnson (1987) for an elementary treatment. Another approach to treating Equations (2.2) is to replace them with the Webster's Horn equation. Both the standard uncurved (see e.g. Chiba and Kajiyama 1941, Fant 1960, Flanagan 1972) and the curved (Lukkari and Malinen 2011b;a) have been used by Aalto (2009) and Aalto et al. (2009) to analyse the same anatomical setup as discussed below. The results are compared under Section 3.3.

In discretising (2.6) with FEM, we obtain three $n \times n$ matrices. They correspond to the three terms in Equation (2.6) in respective order from left to right: $\boldsymbol{K}$ is the stiffness matrix, $\boldsymbol{P}$ represents the glottis boundary condition, and $\boldsymbol{M}$ is the mass matrix. To find the computational approximations for the formants, we have to solve the following linear algebra problem: find all complex numbers λ and corresponding nonzero vectors $\mathbf{x}(\lambda)$ such that

$$\lambda^2 \boldsymbol{K}\mathbf{x}(\lambda) + \lambda c \boldsymbol{P}\mathbf{x}(\lambda) + c^2 \boldsymbol{M}\mathbf{x}(\lambda) = 0 \tag{3.1}$$

With some manipulations (see e.g. Saad 1992), Equation (3.1) can be

written as the standard eigenvalue problem

$$\boldsymbol{A}\mathbf{y}(\lambda) = \lambda \boldsymbol{B}\mathbf{y}(\lambda), \tag{3.2}$$

where $\boldsymbol{A} = \begin{bmatrix} -c\boldsymbol{P} & -c^2\boldsymbol{M} \\ \boldsymbol{I} & 0 \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{K} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}$, and $\mathbf{y}(\lambda) = \begin{bmatrix} \lambda\mathbf{x}(\lambda) \\ \mathbf{x}(\lambda) \end{bmatrix}$. The numbers λ are good approximations of the λ's appearing in Equations (2.5), provided that the number n of elements is high enough and that the computational mesh does not contain any badly shaped elements. In principle, the lowest formants F1, F2, ..., correspond to the numbers λ in the order of increasing imaginary part. In practice, some of the simulated resonance modes might not be excitable from the glottis. This situation arises only with more complex VT configurations (in particular if they include the nasal tract), and therefore is not a problem in case.

3.1.1 Simulation practicalities

The speed of sound is set at $c = 350\frac{m}{s}$. It is the only material parameter needed for the resonance simulation.

We use piecewise linear shape functions and a tetrahedral mesh of $n = 64254$ elements. It is genarated pseudorandomly to fill the geometry shown in Figure 3.1. The geometrical data was collected by Olov Engwall with MRI from a native male speaker of Swedish while he pronounced a prolonged vowel [øː] in supine position.

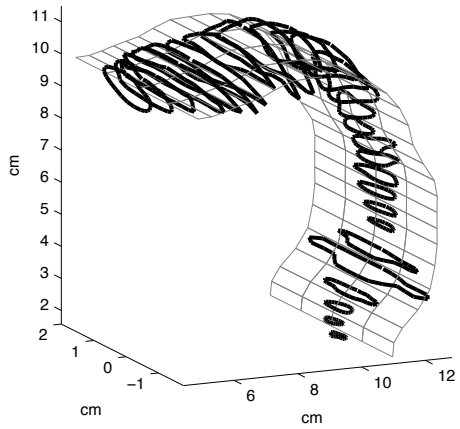


Figure 3.1: The human vocal tract represented by 29 cross-sectional outlines (bold lines). The lips are the last outline on the left and the glottis the one at the bottom on the right. The picture also shows the curving plane (thin lines) along which the Figures 3.3 and 3.2 are plotted.

3.2 Simulated formants

We solved Equation (3.2) in MATLAB environment. The formants F1 to F4 that we obtained are shown in Table 3.1 and Figures 3.2 and 3.3. Table 3.1 also lists formant values obtained from a vocal tract model based on Webster's horn Equation (Aalto 2009, Aalto et al. 2009).

These computed formants are roughly $3\frac{1}{2}$ semitones too high compared to the measured values. We will discuss the physical background of this discrepancy below in Section 3.3. The row labeled "Scaled" in Table 3.1 shows the computed formants multiplied by 0.817, which corresponds to a difference of $3\frac{1}{2}$ semitones. Mainly this observation motivated our work on data acquisition, see Chapters 4 and 5.

Table 3.1: Formants for [ø:] in Hz, from the 3-D wave equation (computed and scaled), from Webster's equation in an uncurved and a curved tube by (Aalto 2009, Aalto et al. 2009) and formants measured by Engwall and Badin (1999)

	F1	F2	F3	F4
Computed	680	1350	2710	3790
Measured	500	1060	2480	3240
Scaled	560	1110	2220	3100
Webster, uncurved	660	1350	2680	3760
Webster, curved	640	1320	2640	3710

We also obtained the resonance modes p_λ – see Equations (2.5) – corresponding to the formants F1-F4. They are computed as linear combinations of the element basis functions, using the components of $\mathbf{x}(\lambda)$ as weights. The perturbation pressures p_λ are not given in any absolute scale here. Rather, they have been normalised so that the maximum deviation from the static pressure p_0 is either 1 or -1. Figure 3.3 shows isobars for the particularly interesting fourth mode. Figure 3.2 shows the pressure distributions of the modes. Figures 3.3 and 3.2 are plotted along a cross-sagittal mid-line plane which is shown in Figure 3.1. Figure 3.3 shows the beginnings of a weak cross-mode resonance related to F4. This supports the prediction that cross-modal resonance becomes a significant phenomenon from around 3 kHz. Also, this makes it necessary to use a full 3D wave equation model even for male speech, if high fidelity results are desired.

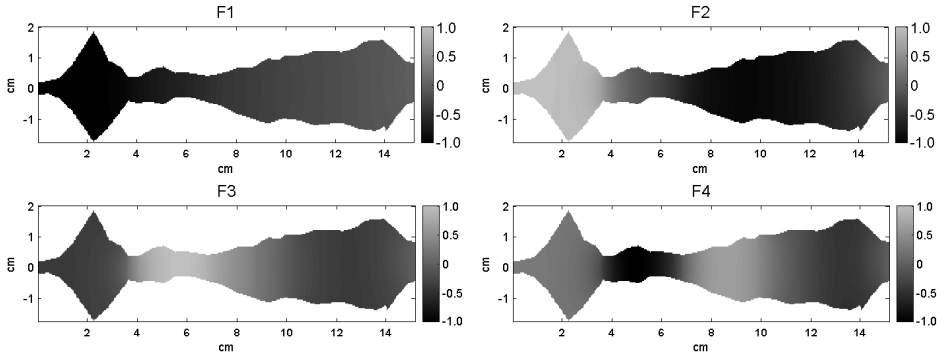


Figure 3.2: Pressure distributions for F1-F4 along a mid-line cut. The mouth is on the right.

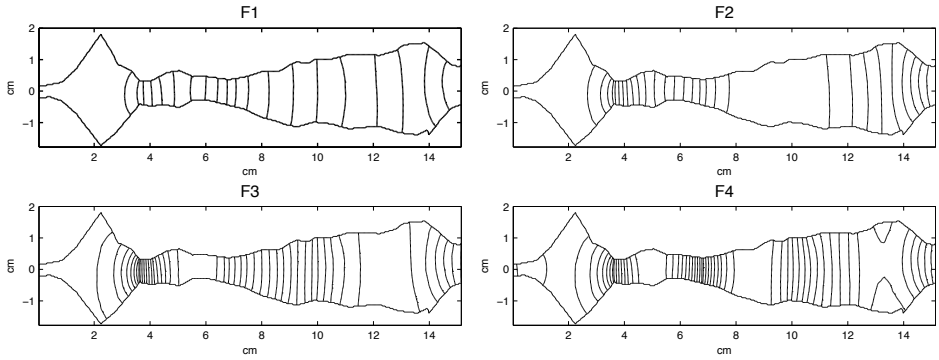


Figure 3.3: Isobars of the pressure distributions for F1-F4 along a mid-line cut. A weak cross-mode can be seen in the mouth in F4.

3.3 Accuracy of the simulation

The vowels from Engwall and Badin (1999: Table 4), together with the computed and scaled $[\text{ø:}]_{c,s}$ from Table 3.1, are plotted in the (F2, F1)-plane in Figure 3.4. Clearly, $[\text{ø:}]_{c,s}$ is closer to the measured $[\text{ø:}]$ than to any other measured vowel, *except* possibly $[\text{ɑ:}]$. To further clarify the situation, let us consider the formants F1 to F4 for $[\text{ø:}]_{c,s}$, $[\text{ø:}]$, and $[\text{ɑ:}]$ as vectors (with values in Hz): $[\text{ø:}]_{c,s} = (560, 1110, 2220, 3100)$, $[\text{ø:}] = (500, 1060, 2480, 3240)$, and $[\text{ɑ:}] = (560, 940, 2740, 3240)$. Then the Euclidean distance between $[\text{ø:}]_{c,s}$ and $[\text{ø:}]$ is 310, but the distance between $[\text{ø:}]_{c,s}$ and $[\text{ɑ:}]$ is significantly larger, equalling 570. This difference is explained by F3, since the fourth formants are almost the same.

Thus, the *first two* formants classify the scaled, computed vowel $[\text{ø:}]_{c,s}$ almost correctly. Moreover, if we look at *all four* available formants, even the remaining ambiguity disappears.

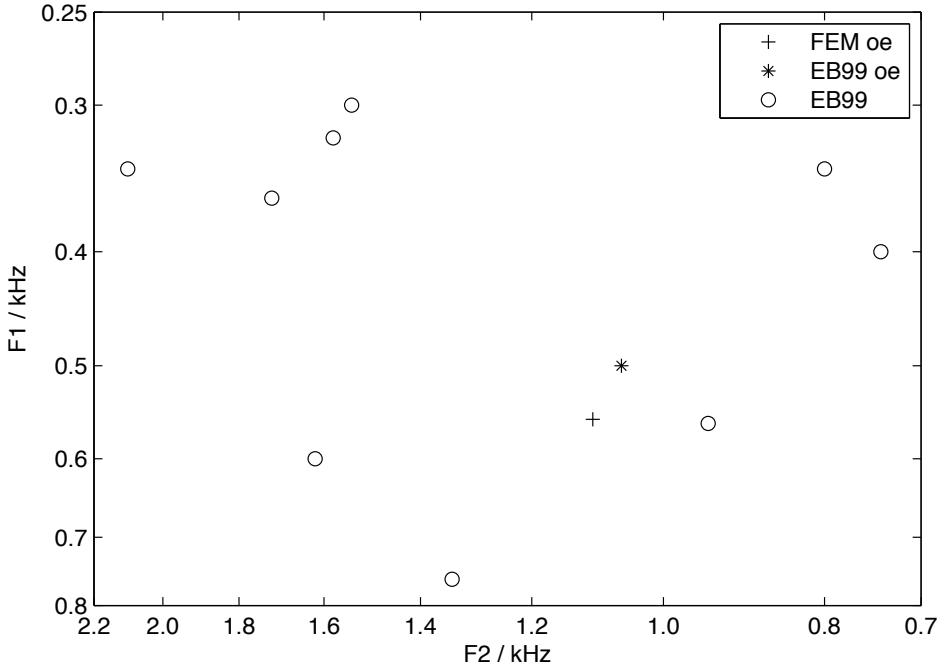


Figure 3.4: Vowels in the (F2,F1)-plane. *FEM oe* (+) is the scaled, computed [øː], *EB99 oe* (*) is the measured [øː] and *EB99* (o) are other measured vowels. (EB99 denotes Engwall and Badin (1999).)

There are, however, other considerations to take into account. First, this is only one geometry and one sound sample and therefore a poor basis for any generalisations. Second, we have here been able to compare only the imaginary parts of the simulated formants to the measured formant frequencies. For a more comprehensive validation also the real parts should be compered to their measured counterparts, i.e., the bandwidths of the formants. Third, the geometry and the sound of the data set used in this experiment were recorded on different occasions and do not necessarily match each other very closely.

The discrepancy of $3\frac{1}{2}$ semitones between all models (wave equation, Webster's horn equation, etc) with respect to the measured data could easily be removed by model tuning. However, we want to construct the models that correspond to the real world phenomena as closely as possible. Thus, it is better to obtain a comprehensive, high quality data set – paying special attention to the match between the acoustic and anatomic data – and base any models on this set.

Chapter 4

Experimental design

As mentioned in the introduction, for the moment our project focuses on vowel production. When considering vowel sounds, the most relevant information are the frequencies of the first two or three formants. When considering vowel articulation, the most relevant information is the overall geometry of the VT, see e.g. Helmholtz (1863), Chiba and Kajiyama (1941), Fant (1960).

As it turned out, our simulation method requires more detailed data than any that we are aware to be freely available. The data of our explorative simulation experiment proved to have three shortcomings which should be remedied. First, it is not possible to generalise from just one data point – i.e. one measured vowel. Second, to validate the model it is necessary to compare the time-frequency structure of the simulated resonances against measured data. Accordingly, we need formant bandwiths in addition to formant frequencies. Third, to maximise the reliability of the validation, the sound and geometrical data should constitute a matching pair – i.e. they must be recorded simultanously.

This Chapter takes a closer look at the requirements for the data and the way we implement them in the measurements. It follows partially the reports by Aalto et al. (2011a;b).

4.1 Requirements

A clear requirement in all speech studies is acquiring the data with minimal risk to the test subject and (in most) on as naturally produced speech as possible. In addition, there are technical requirements imposed by our choice of simulation methods and other factors.

The first goal in combination with the fact that we need to image the

whole VT – and not for example only the tongue – limits effectively our choice of imaging methods to MRI. The only other method, which is capable of producing 3D images of the whole VT, is X-ray based computed tomography (CT). It exposes the subject to ionising radiation, which might be acceptable in studies which require only a small amount of data. However, in this case repeated measurements of a great number of utterances are likely to be needed at some point, and thus MRI has to be our method of choice.

In detail, the design requirements are as follows. First, in general for the experimental setting:

- The setting must not compromise safety of the test subject.
- A suitable male subject, i.e., somebody who is safely and without compromising image quality able to act as a subject of an MRI study, preferrably a trained phonetician and singer.
- The setting has to be as comfortable as possible to facilitate natural speech production.
- Access to clean speech signal in real time.
- The phonetic materials must support the implementation and validation of the computational model.

Second, specifically for the data:

- Anatomical data with a spatial resolution of about 1mm.
- As fast an imaging sequence as possible.
- The fundamental frequency f_0 before, after and *during* the MR imaging sequence.
- As many of the formants (F1, F2, F3, F4, ...) as possible,
- and their bandwiths before, after and *during* the MR imaging sequence.

There are inherent conflicts between these requirements. For example, a fast imaging sequence usually does not have a good spatial resolution. Also, it is difficult to get a clean sample of speech while the MRI machine is running as it produces strong acoustic noise. Furthermore, safety and comfort issues concerning the human test subject restrict the design of all the equipment to be used. Overall, these considerations lead to a need to strike a balance between the different aspects of the experimental setting. Our approach to tackling these challenges is detailed in the following sections and Chapter 5.

Engwall (2000; 2003; 2006) has shown that the supine position impedes the naturalness of speech. However, this problem is more or less inescapable as computed tomography also requires a supine position.

4.2 Experimental setting

Figure 4.1: Test subject and equipment under preparation for the first recordings. Dr. Tech. Jarmo Malinen (left) is placing the sound collector on the MR registering coil over the face of the test subject (middle) while Dr. Jani Saunavaara (right) is observing.

The MRI room is a quite challenging sound recording environment. The challenges and solutions of sound recording are detailed in Chapter 5. When using MRI the subject will lie in a supine postion inside the bore of the main coil of the imaging machine. Figure 4.1 shows the subject being prepared for insertion into the machine. In addition to our recording equipment in front of his face, the subject wears earphones. They are intended for dampening the acoustic noise arriving at the subject's ears and to play, e.g., music while the examination is on the way. We have recruited the earphones for providing verbal and automated instructions for the test subject.

When lying inside the machine, the subject's vision is even more limited then during normal MRI studies because the sound collector lies in front of his face. The bore is fairly small, and during an imaging sequence, the machine produces strong acoustic noise. Taking these factors into account, it would be preferrable to keep the recording sessions as short as possible. However, limited by the practicality of setting up the equipment

a recording session at the pilot stage lasts for a fairly long time: on the order of 0.5-1 hours.

4.3 Phonetic materials and subject

Humans are able to produce vowel sounds which are perceived as the same vowel with significantly different vocal tract shapes. These vowels share the main acoustic characteristics, and for this reason they present a challenge for data acquisition. A subject may sound as if he is producing an unchanging vowel while his vocal tract is actually changing its shape all the time to accommodate changes in the rest of the articulatory system.

There are three main sources of movement artefacts in sustained vowel production: 1. adjustments due to gravity (Engwall 2003, Stone et al. 2007), 2. changes in larynx position due to altering the fundamental frequency, and 3. changes caused by contraction of the thorax during a long exhalation.

Accordingly, we had the subject produce vowels at two constant fundamental frequencies: 110 and 137.5 Hz (i.e., notes A2 and C#3). The productions were imaged with two different MRI techniques: a stationary 3D sequence and a dynamic 2D mid-sagittal sequence. Thus, we gained data on the effect of larynx position as well as the changes caused by a contracting thorax.

We gathered data on the Finnish vowels [ɑ, e, i, o, u, y, æ, ø] with $f_0 = 110, 137.5$ Hz and [æ-ɑ] glides. The materials were produced by a 30-year-old healthy male subject, who has a background in phonetics and singing and is a native speaker of Finnish.

Before the experiment, the subject was given a description of what he would be asked to do next. The experiment was started when the subject indicated that he was ready. First, the subject heard a sinusoidal cue signal that gave him a count-down for starting the utterance at the right time as well as the desired pitch, i.e., the level of f_0.

A typical sound sample, including the cue signal, is represented in Figure 4.2. The MRI machine was operated so that a 500 ms "pure sample" of stabilized utterance could be obtained immediately before and right after the MRI noise interval.

After each experiment, the image data was inspected visually and the subject gave his comments. The sound sample was listened to by a trained phonetician in the control room during the whole imaging sequence, and unsuccessful utterances were usually detected immediately. Particular attention was paid to the phonation type and nasality.

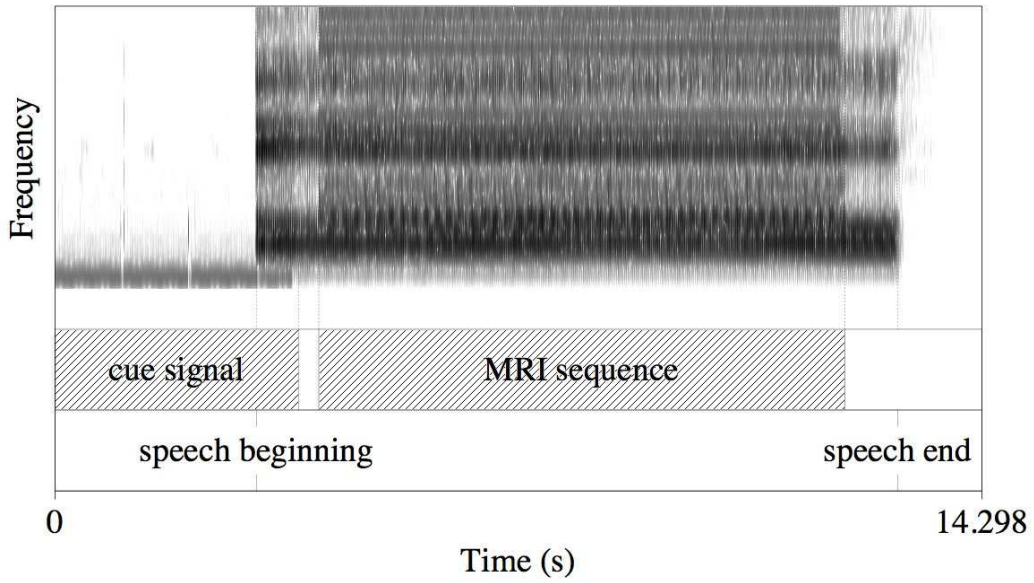


Figure 4.2: A spectrogram showing a full sound recording. From left to right: cue signal for the subject ($\approx$ 3.5 s from the signal onset, overlapping speech for the last $\approx$ 500 ms); the clean speech sample ($\approx$ 500 ms); the speech and the imaging noise ($\approx$ 8 s); and the clean speech sample ($\approx$ 500 ms). Hatched background indicates interference the cue signal and the MRI noise cause in the speech recording. The time windows when a clean speech sample is available are indicated by the absence of hatching.

4.4 MRI sequence

We carried out the imaging using a Siemens Magnetom Avanto 1.5 T system (Siemens Medical Solutions, Erlangen, Germany) at the Medical Imaging Centre of Southwest Finland. This scanner has a maximum gradient field strenght of 33 mT/m (x,y,z directions) and a maximum slew rate of 125 T/m/s.

We combined a 12-element Head Matrix Coil with a 4-element Neck Matrix Coil to be able to image the whole upper airway. With this configuration, we could use Generalized Auto-calibrating Partially Parallel Acquisition (GRAPPA) to reduce scan times (Griswold et al. 2002). We used the technique in all of the scans with acceleration factor 2.

After comparison with other sequencies, we used 3D VIBE (Volumetric Interpolated Breath-hold Examination) (Rofsky et al. 1999) for the 3D scans[1]. This sequence is an ultra-fast gradient echo sequence with

[1]The sequence was originally developed for 3D imaging of the abdominal area.

isotropic resolution. Its k-space scan is typically performed asymmetrically. This reduces the number of phase encoding steps in the slice selection direction which leads to faster scan times. We optimised the sequence parameters to minimize acquisition time. This lead to choosing 1.8 mm voxel size for the VIBE scans (called *VIBE 1.8* for short) as well as a dynamic sequence with a frame rate of 5.5 images per second and a variable length in time (*dynamic* for short). The parameters of the sequences are detailed in Table 4.1.

Table 4.1: The MRI sequencies and their parameters. (The abbreviations are TR = time of repetition, TE = echo time, FA = flip angle, BW = receiver bandwidth, and FOV = field of view.)

Sequence	Duration	TR	TE	FA	BW
3D VIBE	7.6 s	3.36 ms	1.19 ms	6°	600 Hz/pixel
Dynamic	n/a	178 ms	1.4 ms	6°	651 Hz/pixel

Sequence	FOV	matrix	Others
3D VIBE	230 mm	128x128	44 slices, slab thickness 79.2 mm
Dynamic	230 mm	120x160	slice thickness 10 mm

4.5 Sound measurements on long phonations

The results from the first MRI recordings made it clear that we need a deeper understanding of what a long vowel production actually is to produce the best possible data. Especially, we need to understand the characteristics and factors affecting a long production's stability.

To achieve this goal, we performed sound measurements in an anechoic chamber at the Aalto University Department of Signal Processing and Acoustics in Otaniemi. We used the same test subject, body position, and phonetic materials as in the MRI study above. The acoustic environment of the MR imaging situation was not modeled in any way; i.e., the subject produced the samples in a quiet room without any reflecting surfaces near him (apart from the board he was lying on). Figure 4.3 shows the measurement situation.

The phonations were of two categories: 'long' ranging from 19.24 s to 23.04 s, and 'short' ranging from 9.99 s to 15.27 s in duration. The samples were recorded with a high quality omnidirectional condencer microphone

AKG C392 and digitised using a high quality A/D converter DigiDesign M-Box.

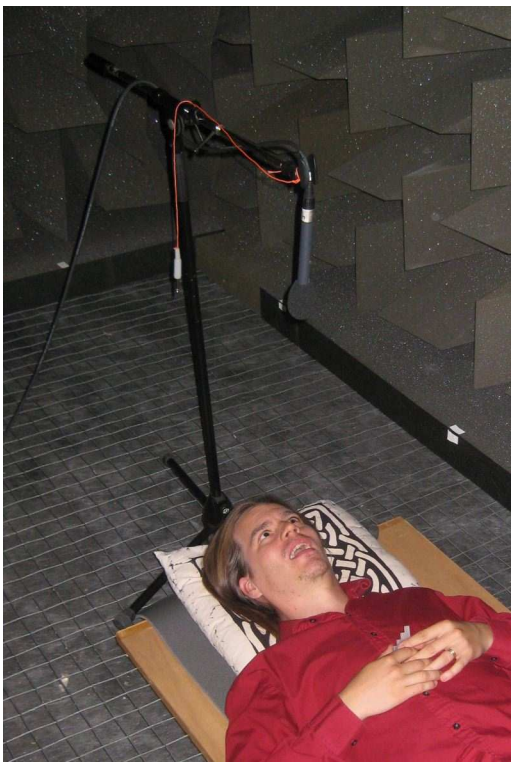


Figure 4.3: The measurement situation of the long phonations in the anechoic chamber.

Chapter 5

Sound recording arrangement

The noise cancellation can be ideally understood as separating a target signal (i.e., the speech) from a cylindrically symmetric noise source (i.e., the environment) while paying attention to the technical complications described below. In addition, there are acoustic complications described in Sections 5.1 and 5.3.

There is acoustic noise of about 90 dB(SPL) during the imaging sequence. The noise is mainly generated by the gradient magnetic field of the MR imaging sequences. The noise is associated with the rapid alterations of currents in the gradient coils that introduce vibrations to the gradient coils. The noise is produced when the gradient coils move against their mountings. A lesser constant background noise is produced by the helium pump of the MRI scanner's cryogenic system.

The noise characteristics vary with imaging parameters, which affect gradient output, such as rise time and amplitude. The intensity of the noise tends to be inversely proportional to section thickness, field of view, repetition time, and echo time. The noise is also dependent on the type of the MR system, presence and size of a patient/subject, room acoustics, etc. The acoustic half life of

To make matters worse for noise cancellation, multi-way propagation is a significant phenomenon inside the bore of the MR imaging coil. Echoes of the noise arrive at the sound collector with different delays.

A Siemens Magnetom Avanto 1.5 T MRI machine produces a static 1.5 T magnetic field, and an imaging sequence produces an electromagnetic field at 64 MHz with a peak power of several kW. Because of safety and image quality considerations, no metal or electronics can be taken near the test subject.

Because of the magnetic field, only negligible amounts of ferromagnetic material may be used in the experimental apparatus inside the MRI

room. None at all is allowed in the sound collector inside the MRI main coil. All electronics inside the MRI room have to be shielded against overvoltage and radio frequencies. Closed loops in all conducting material must be strictly avoided.

It is necessary to distinguish between sound artefacts generated by the recording equipment and the actual frequency distribution of the subject's speech. This is possible only by knowing the frequency response of the recording equipment. With this knowledge, it becomes possible to either correct the imperfections by modifying the equipment or by signal post-processing.

The sound recording arrangement is reported in Lukkari et al. (2007), Malinen and Palo (2009), and Aalto et al. (2011a).

5.1 Principle of noise cancellation

The recording setup is based on the principle of a differential microphone operating in dipole configuration. To accommodate the above mentioned technical complications, the microphones themselves are located some distance from the subject, and the sound samples are collected by a passive sound collector and transported to the microphones via acoustic waveguides. Now, let us take a closer look at how the system works in principle.

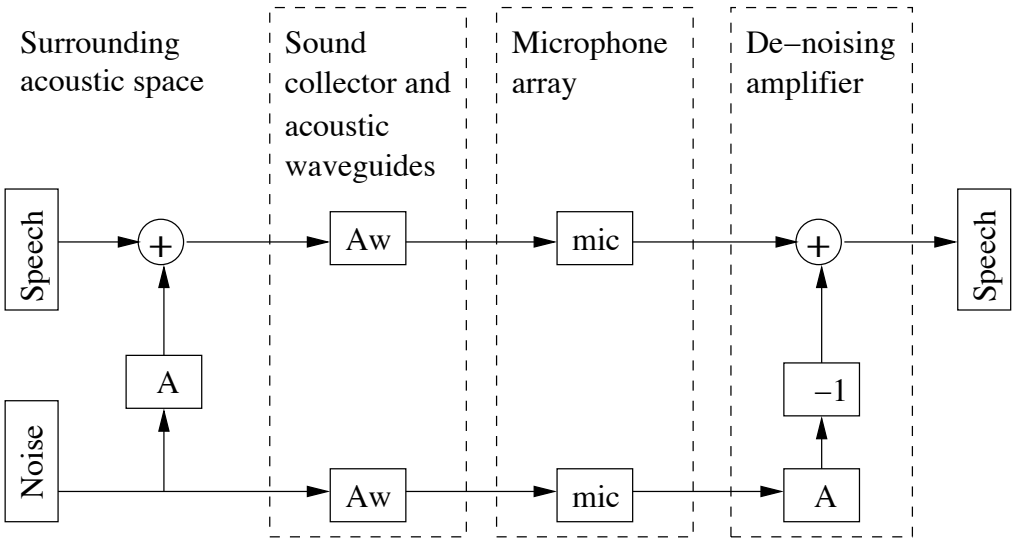


Figure 5.1: System level schematic of the sound recording setup

Figure 5.1 shows a simplified system level view of the sound record-

ing setup. In an ideal situation, the outputs from the sound collector would be:

$$\begin{cases} x + A(y) \\ y, \end{cases}$$

where x is the speech signal, y the noise signal caused by the MR imaging sequence, and A a frequency dependent level difference between the noise in the speech channel and the noise in the noise channel. The speech and noise samples will then be filtered by the acoustic waveguides

$$\begin{cases} A_w(x + A(y)) \\ A_w(y), \end{cases}$$

where A_w is the linear filter function of the wave guides. With the de-noising amplifier we correct the level difference between the noise signals and deduct the weighted noise sample from the contaminated speech signal to arrive at an estimate of the speech

$$\begin{aligned} \hat{x} &= A_w(x + A(y)) - \hat{A}(A_w(y)) \\ &= A_w(x) + A_w(A(y)) - \hat{A}(A_w(y)) \\ &\approx A_w(x), \end{aligned} \tag{5.1}$$

where $\hat{A}$ is the de-noising amplifier's gain, which has been adjusted so that $A_w(A(y)) - \hat{A}(A_w(y)) \approx 0$. The spectral distortion effect of the acoustic wave guides remains to be removed with post-processing.

The treatment above is a good first approximation, and it excludes many of the imperfections of the situation. Two of the most important ones are frequency dependent phase difference between the channels and multi-way propagation. Neither of these affects the system on low frequencies, but around 1 kHz they become increasingly significant phenomena. The phase difference is caused by the distance between the receiver horns of the two channels.

Multi-way propagation becomes significant for the same reason, but has a more complex nature. While the phase difference is easy to anticipate by calculations, multi-way propagation is caused by different surfaces reflecting acoustic signals inside the MR scanner – especially the noise – and causing several delayed copies of it to arrive at the sound collector, out of phase with each other. Since the internal geometry of the scanner is somewhat complex, the resulting noise signal has an inconvenient, position dependent structure. Obviously, this would not be a problem if both channels received the same copies of noise at the same time. Since the

channels have to be at a distance from each other, problems do arise from multi-way propagation.

5.2 Physical recording setup

Unfortunately, it is nigh impossible to take pictures of the recording setup in the MRI room because of the magnetic field. The setup had to be reproduced in an anechoic chamber for the frequency response measurements, and our pictures are from there.

So, to get an impression of the physical setup let us look at Figure 5.2 which shows the laboratory arrangements in the anechoic chamber. In the Figure the setup can be seen from left to right as follows: noise reflector, reference microphone probe, sound collector, and face model. A reference sound source (not shown in this picture) is situated at the face model's mouth. The face model is part of the response measurement setup and it is situated at the same relative position to the sound collector as the test subject is in the MRI scanner.

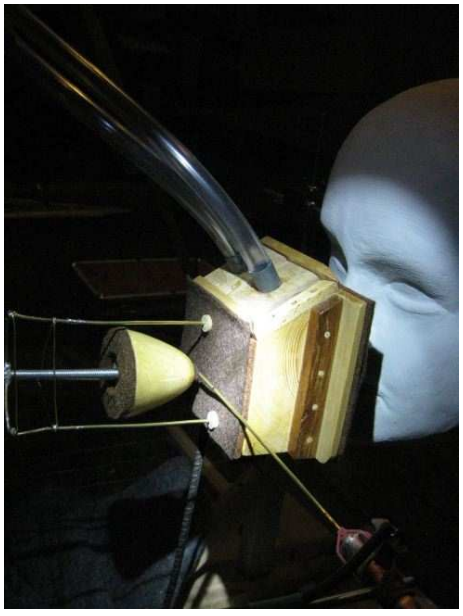


Figure 5.2: Laboratory arrangements for the sound collector measurement in the anechoic chamber.

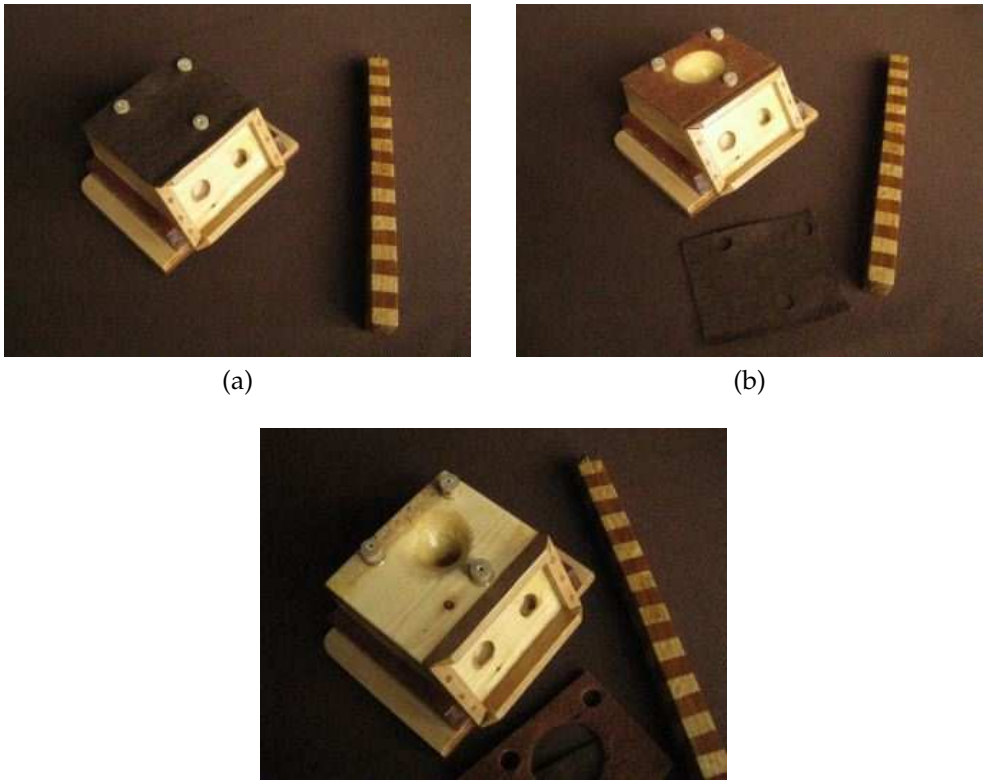
5.2.1 Sound collector

We use a two-channel sound collector, see Figures 5.4 and 5.5. It is completely metal free, passive, non-microphonic, and does not have any moving parts. One channel is reserved for the speech sample and the other for the noise sample. The dimensions of the collector are small compared to the lower formant wavelengths, and the collector fits on top of the MRI head coil.

(a) (b) (c)

Figure 5.3: (a) From left to right: the face model, a reference microphone and the sound collector (b) Prototypes for a paraboloid reflector (c) One of the paraboloid reflectors suspended from a prototype support structure.

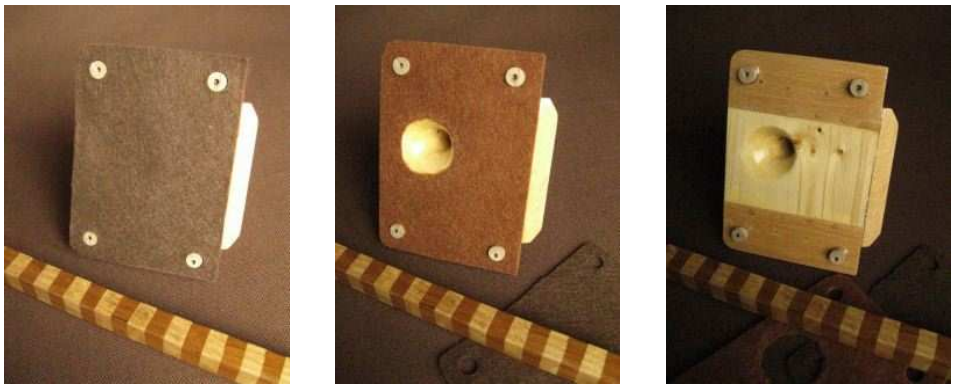
The test subject's face affects significantly the acoustic field around the sound collector. To match the acoustic properties of the two channels, we use a paraboloid-shaped reflector on top of the sound collector at the opening of the noise channel. More importantly, we avoid taking a noise sample from a small point on the ceiling of the MRI coil's bore by shadowing the centre of the noise channel horn's directional cone with the paraboloid reflector. Three prototype reflectors and the position of the reflector in relation to the sound collector are shown in Figures 5.3b and 5.3c. During data acquisition in MRI, we use a suction cup to attach the reflector to the ceiling of the MRI bore. In addition to the reflector, we fine tune the acoustics of the sound collector with layers of damping material on both of its surfaces. These layers are shown in Figures 5.4 and 5.5. The outer layer dampens the longitudinal resonances of the acoustic wave guides and prevents exhalation air flow from causing artefacts in the recorded sound.



(a) (b)

(c)

Figure 5.4: The sound collector (next to a centimeter scale) seen from above (noise channel side) with 2 (a), 1 (b), and without (c) layers of damping material.



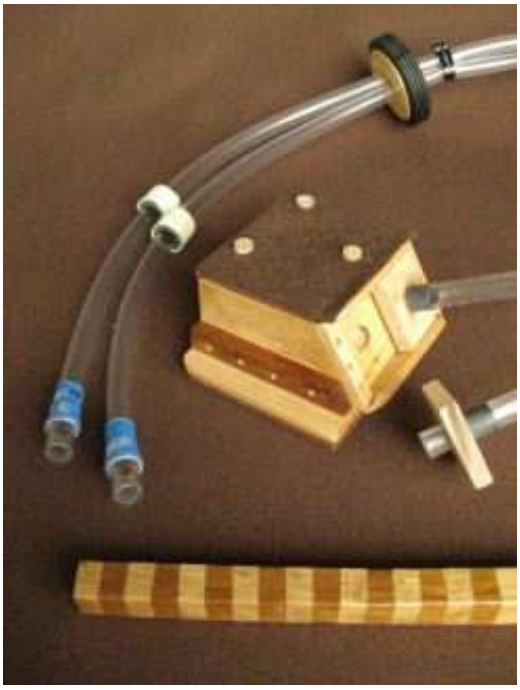
(a) (b) (c)

Figure 5.5: The sound collector (next to a centimeter scale) seen from below (speech channel side) with 2 (a), 1 (b), and without (c) layers of damping material.

5.2.2 Acoustic wave guides

The sound signals are transmitted to a microphone assembly by acoustic wave guides, see Figure 5.6a. The medium in the collector and the wave guides is air. Sound transmission in the wave guide walls appears to be negligible, by an oscilloscope measurement of transmission delays.

The waveguides are constructed from soft PVC tube of inner diameter 9 mm. The length of each wave guide is 3000 ± 1 mm. They are suspended pairwise from statives so as to cancel out external disturbances. A stative is shown in Figure 5.6b. The statives are free of magnetic materials and the stative used inside the MRI main coil is free of metal. The waveguides are attached to the sound collector and the microphone array at the opposite ends. These attachments can be seen in Figures 5.6a and 5.8, respectively.



(a)

(b)

Figure 5.6: (a) The acoustic wave guides (next to a centimeter scale) with one attached to the sound collector and the other showing the details of the attachment mechanism. (b) A stative used in suspending the acoustic waveguides

5.2.3 Shielded microphone array

The microphone assembly is enclosed in a Faraday cage as shown in Figures 5.7 and 5.8. The cage is made of 6 mm aluminium plate, which is thick enough not to buckle or resonate. Damping material is used inside the cage. The acoustic wave guides are brought into the cage through electromagnetic waveguides, designed to be opaque at frequences between 10–100 MHz. More importantly, the electromagnetic waveguides allow acoustic insulation to be implementd for the acoustic waveguides' the insertion points which would otherwise be likely to leak. As it is, the Faraday cage is practically completely quit when the lid is in place on top of it.

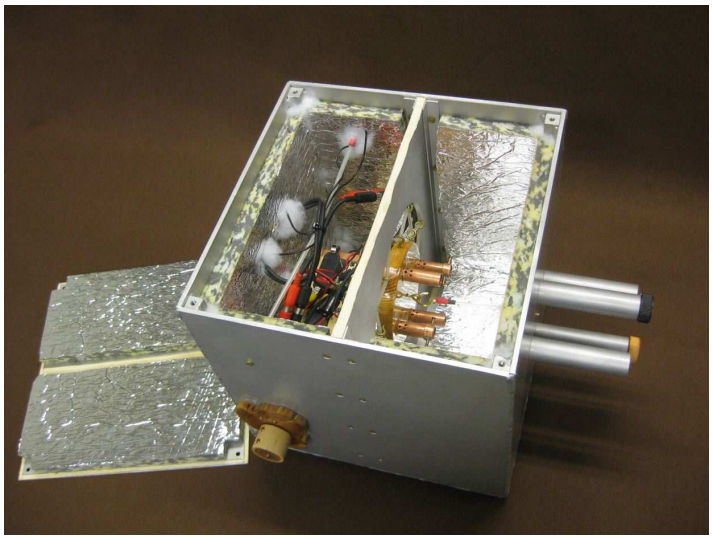


Figure 5.7: The microphone array inside the Faraday cage.

The microphone assembly (see Figure 5.8) consists of four Panasonic WM-62 condenser microphones (with sensitivity -45 $\pm$ 4 dB re 1 V/Pa at 1 kHz, $\varnothing$ 9 mm) and a 5 V regulated, battery-driven power source for them. The nominal frequency response of the microphones, as given by the manufacturer's data sheet, is essentially flat in the frequency range of interest. By a superficial measurement, sensitivities and frequency responses of such microphone units do not seem to differ from each other (or the nominal values) significantly, and hence we omitted more detailed measurements.

The microphones are embedded into a wooden plate that is acoustically and electrically isolated from the walls of the Faraday cage. The sound waves enter the microphones through simple, adjustable acous-

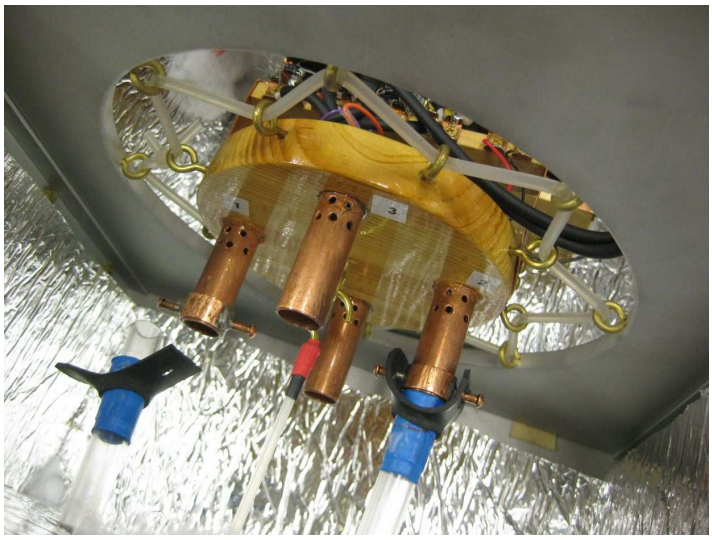


Figure 5.8: Microphone array

tic impedance matchings, see the lower right corner in Figure 5.8. These matchings consist of holes with ∅ 2 mm in the walls of the copper tubes. The matchings partially supress the longitudinal resonances of the wave guide (see Figure 5.13). An energy dissipation of several dB's is measured in the frequency response of the system because of these holes. Since the matching consists of both open and closed partial terminations of the wave guide, the residual reflection takes place both with and without phase inversion. This corresponds exactly to the number of measured peaks in Figure 5.13 in the 0.5-2 kHz range.

The signals are transmitted from the MRI room by two microphone cables (Tasker C116 4x0.14-26AWG); two channels in each. All cable endings are shielded against overvoltages by diodes. Since only two channels are used by the current sound collector, the remaining microphones are a reserve.

5.2.4 De-noising amplifier

As mentioned in Chapter 4, the test subject needs to hear the de-noised signal in real time. Hence, we implement the de-noising system as an analog device. It is a summing amplifier (see Figure 5.9) with one direct channel (for the signal) and three adjustable, inverted channels (for subtracting up to three noise signals). Before recording, the summing coefficients are adjusted manually by listening to the output. The main components of the device are six operation amplifiers of type LM741, and its input impedance

is 3 kΩ. The inputs have been RF- and overvoltage shielded by chokes and diodes.

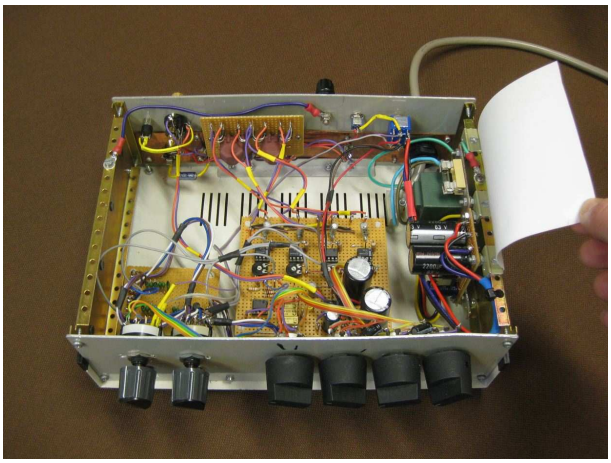


Figure 5.9: De-noising amplifier

5.3 Response measurements

The face model has a hole running through it to its mouth, allowing us to place the pointwise sound source at the mouth as shown in Figure 5.10. In the Figure there are (starting from the left top corner) the acoustic waveguides (going out of the picture), the sound collector with a reflector paraboloid in place, beneath them the microphone probe, and continuing to the right: the face model and on the far right the pointwise sound source, whose acoustic chamber is the rectangular block. In the setup shown in Figure 5.10 we use an additional sound source to measure the response of the noise channel. On the other hand, if the microphone probe were moved to the center of the speech channel, we could use the pointwise sound source to produce a calibration chirp. In this manner we can measure frequency responses that are shown in Figure 5.13 where the microphone probe has been used to sample the sound received by the system.

The measured frequency response of the sound recording system is a combination of the effects of the acoustic waveguides, the sound collector and the subject's face. The last is demonstrated by the two curves in Figure 5.11. Both curves show two distinct regions. Under 1.7 kHz we see both exhibiting a rising peak pattern which corresponds to the multiples of half and quarter wave resonances of the acoustic waveguides.

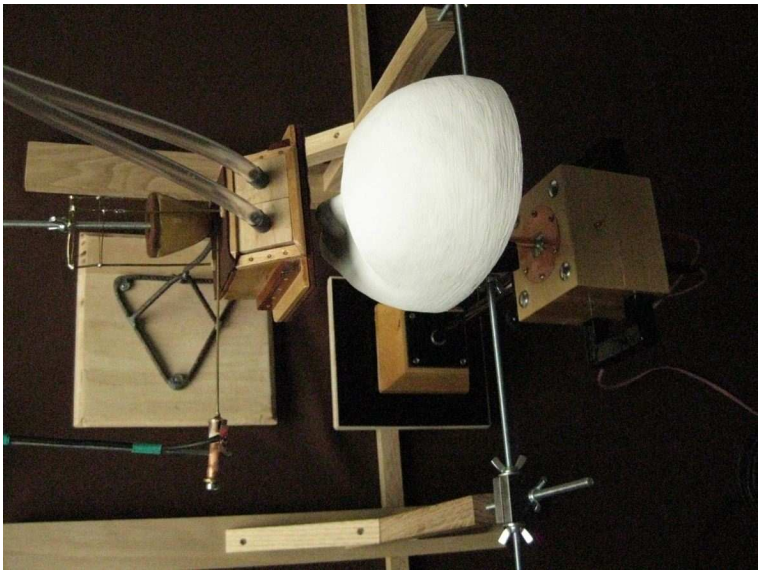


Figure 5.10: A mock up of a noise channel frequency response measurement.

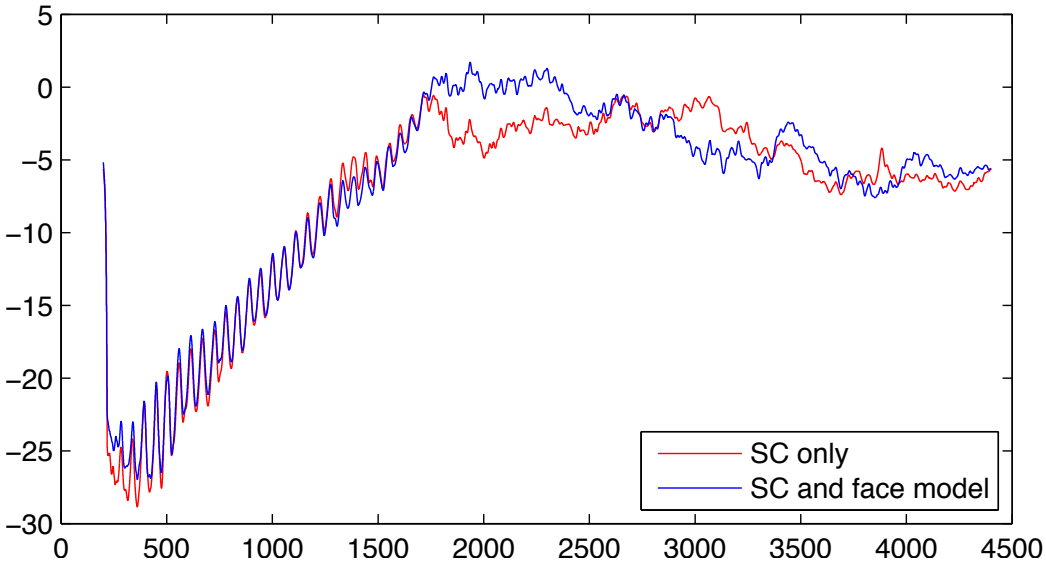


Figure 5.11: Frequency responses of the speech channel with (red) and without (blue) the face model measured with the arrangement in Figure 5.3a. (X-axis corresponds to frequency (Hz) and Y-axis to attenuation (dB))

Above 1.7 kHz the curves differ more significantly from each other. This is due to the effect of the face model on the system's acoustic properties and the fact that the waveguide resonances are no longer a significant phenomenon in this region.

Now, the curves in Figure 5.11 were measured without the microphone probe with just a naked micophone as in Figure 5.3a. The microphone was found to be large enough to have a significant effect on the measured frequency response. This is illustrated by Figure 5.12 which shows the difference in the responses which were measured with the reference microphone close to the sound collector and sound source as in Figure 5.3a, and with the microphone removed to a greater distance.

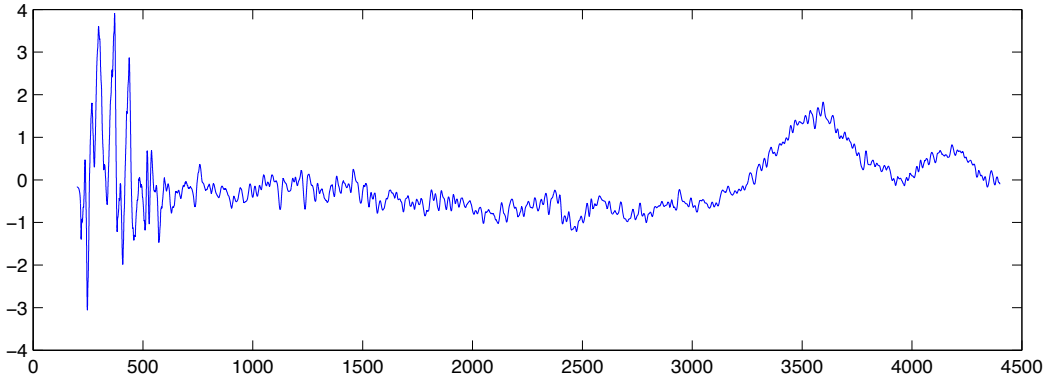


Figure 5.12: The effect of a microphone-sized object on the frequency response of the speech channel. (X-axis corresponds to frequency (Hz) and Y-axis to attenuation (dB))

This response would be flat in an ideal situation. As seen the measurement arrangement is quite sensitive even to the (spatial) volume of teh small microphone. Even if this was not the case, the microphone could not physically fit in to the reference point between the mouth and the sound collector. This observation lead to the design of the microphone probe, which has a negligible effect on the responses measured with it. The frequency response of the probe was measured carefully, and the effect was removed numerically. The directional behaviour of the probe was not measured as the probe was assumed to have point-wise behaviour.

The frequency response of the acoustic wave guide between 0.3–4.4 kHz is given in Figure 5.13. At lower frequencies in Figure 5.13, longitudinal resonances of the wave guide appear. Below 2 kHz, there is $\approx$ 4dB attenuation per octave that can be easily compensated with DSP during post processing. Although Sondhi (1986) shows that the curvature of an acoustic wave guide affects its resonances, in our case the relevant dimen-

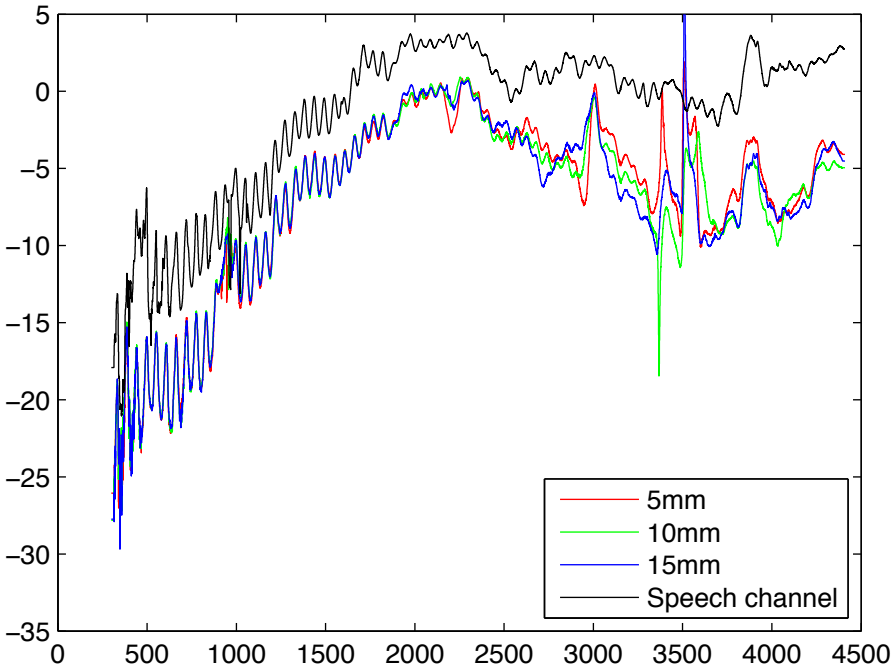


Figure 5.13: Frequency responses of the speech channel (black, raised above the others for clarity) and of the noise channel with three different paraboloid distances (red, green and blue) measured with the arrangement in Figures 5.2 and 5.10.

sions and magnitudes make the effect negligible (Lukkari and Malinen 2011b).

Please note, that the frequency responses of the speech channel in Figures 5.11 and 5.13 are measured in significantly different conditions. In the former, the reference microphone was not fitted with the probe, the system was not used in dipole mode, and the sound collector was not covered with any damping material. In the latter, all of these factors were in effect. Furthermore, in the latter case, the probe's point was positioned at the centre of the sound collector horn, 5 mm above the surface. At this point, the sound pressure was kept constant.

5.4 Computer equipment and digital signal processing

The de-noised signal from the amplifier is digitized using a MacBookPro2,2 computer running MacOSX 10.4.9. The required signal processing and formant extraction is done using MATLAB 7.4, Signal Processing Toolbox, and custom made code. All recordings – whether in MRI, in the anechoic chamber or for response measurements – were sampled with 44100 Hz sample frequency.

There is considerable detail involved in measuring and compensating the responses of the microphone probe, the pointwise sound source and the additional sound source. This involves first measuring a chirp signal. Then generating a new frequency weighted chirp based on the measured signal and iterating until the change from one run to the next is negligible. This is a process of fixed point iteration, implemented with hardware and signals instead of numerical software. Additional complications arise from the behaviour of loudspeaker elements and from the resonances of different parts of the sound sources.

Chapter 6

Measurement results

We have conducted a pilot data acquisition. We used the setup for recording and MRI described in Chapters 4 and 5. Some results are also reported in Aalto et al. (2011a).

As mentioned in Chapter 4, in this thesis we concentrate on the data gathered on the Finnish vowels [ɑ, e, i, o, u, y, æ, ø] with $f_0 = 110$ and 137.5 Hz and [æ-ɑ] glides with $f_0 = 110$ Hz.

6.1 Sound data from MRI

For sound analysis, 150 ms sound samples were chosen by the author from each of the recorded speech samples. That no MRI noise was present in the samples was ensured by listening to the samples and examining spectrograms. In some cases, it was necessary to compromise the quality of the speech sound in order to avoid the MRI noise. The sampling was done both before and after the MR sequence so as to be able to tell whether the sound uttered by the subject had changed significantly during the imaging.

At this stage, no noise cancellation – besides that performed by the analogue de-noising amplifier, (see Figure 5.9) – was applied, and thus analysing the sound recorded during the imaging sequence remains future work.

6.1.1 f_0 and formant extraction

The fundamental frequency (f_0) and the first four formants (F1-F4) were extracted from the 150 ms sound samples with MATLAB. The values of f_0 and (F1-F4) thus obtained are shown in Tables 6.2 and 6.3 located at

the end of this Chapter. The tables also list additional acoustic measures: Sound pressure level difference between the begin and end samples, and formant distance between the samples (see below for the algorithm used in calculating the formant distance). Figure 6.1 shows a F2-F1-plot of the data set.

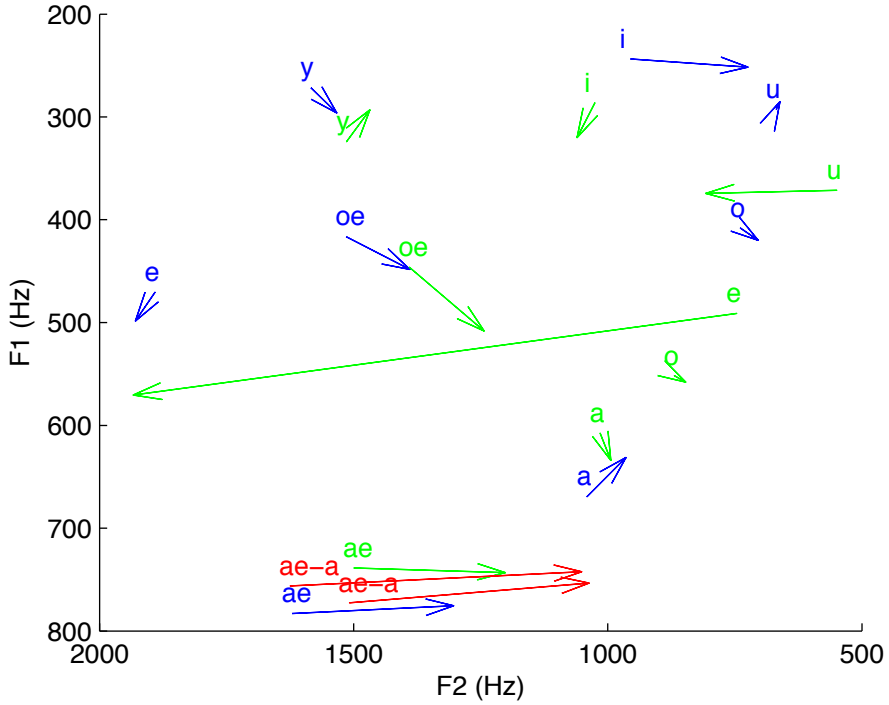


Figure 6.1: Formant shifts of the MRI data set in F2-F1 plane. The arrows point from the begin values to the end values. The blue arrows represent the 110 Hz data, the green arrows the 137.5 Hz data and the red arrows the 110 Hz glide data.

In the F2-F1-plot, the effect of an artefact in the sound data at about 1 kHz can be seen: The [e] with f_0 = 137.5 Hz clearly has an aliased F2 in the begin sample and the [i]s have an aliased F2 in both samples. This artefact is most likely caused by sound reflecting from the inner surfaces of the MRI registering coil. The artefact is also visible in the spectra described below and is discussed in Chapter 7.

The fundamental frequency f_0 was estimated with autocorrelation. The values were constrained to the range 100-150 Hz to avoid octave jumps. The formants were extracted with a linear prediction (LPC) model of 45th degree (Atal and Schroeder 1967). Both of these analyses were per-

formed without compensating for the frequency response of the recording system in any way. The reported formants are the first four lowest roots of the model. In addition, formant distances were calculated between the beginning and end samples. The algorithm (in MATLAB syntax) used is listed below:

```
for i=1:4,
  r(i) = min(abs(formants_b(i)-formants_e));
end
formant_distance = sqrt(sum(r.^2));
```

Here *formants_[b,e]* contain all roots of the LPC model for the beginning and end samples. This is not strictly a Euclidian distance between the vectors formed by the first four roots of the model. Rather than pairing each estimated formant from the begin sample with the corresponding from the end sample, it pairs each begin formant with that of the end formants which is closest to the begin one. This means that it is a good measure of change as it does not penalise cases where the LPC algorithm finds an extra peak between two formants. Figures 6.3 and 6.4 show examples of LPC spectra. See the end of the following section for explanation of the colours used in the Figures. The complete data set is in Appendix A.

6.1.2 Spectral analysis

Spectral analysis was carried out with MATLAB. FFT spectra of the 150 ms speech samples were obtained, compensated, and smoothened. The spectra were compensated with the frequency response of the speech channel of the recording system below 4.4 kHz. The compensation algorithm was a simple dB-scale deduction of the response from the raw FFT spectrum. The response was extrapolated for frequencies below 300 Hz. Figure 6.2 shows the response as it was used in the compensation.

The smoothening was performed with a moving triangle weighted average filter which was applied first forwards and then backwards to obtain a smoothened spectrum with unshifted frequencies. The length of the filter was chosen to correspond roughly to the fundamental frequency of the sample to be processed.

As a parallel approach, we calculated spectrograms. These were used for calculating time average spectra (producing an intensity spectrum rather than a power spectrum). The spectrograms were generated with a 10ms Gaussian sampling window and an overlap of 8.75 ms between adjacent samples.

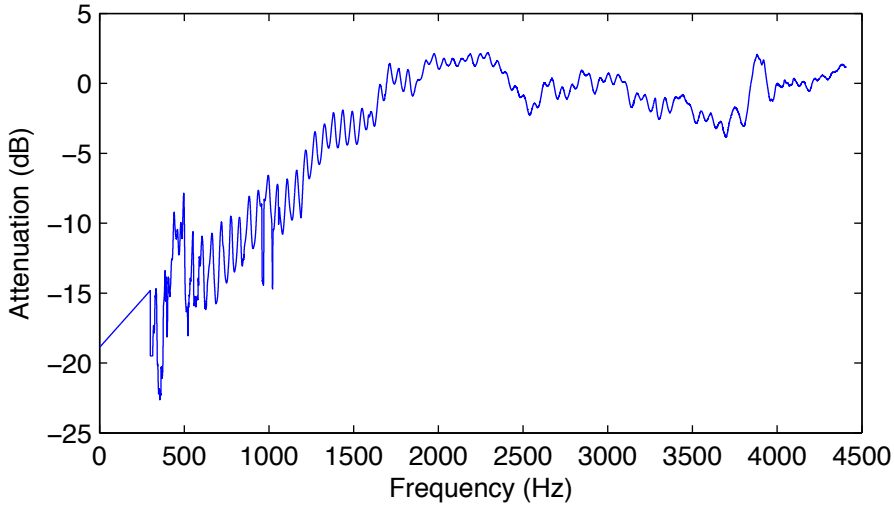


Figure 6.2: The speech channel's frequency response as used in spectral compensation of the smoothened spectra

Examples of the obtained spectra and spectrograms are shown in Figures 6.3 and 6.4; for a complete set, see Appendix A. The color code for the curves is as follows: Blue corresponds to the begin sample, green to the end sample and red to their spectral difference. In the LPC plots, the solid blue vertical lines represent the estimated formants of the begin sample and the dashed green vertical lines those of the end sample. In the spectrum plots, dotted vertical lines simply mark the frequencies 500, 1000, 2000, 4000, and 8000 Hz while the dotted horizontal line marks the 0 dB level.

6.1.3 Acoustic noise data

We recorded also the acoustic noise generated by the imaging sequences we used. The recording was done with the same setup as the speech recordings. The only difference was that the test subject lay silent within the system instead of speaking. Spectra and spectrograms calculated in the same manner as those for the speech samples are shown in Figures 6.5 and 6.6.

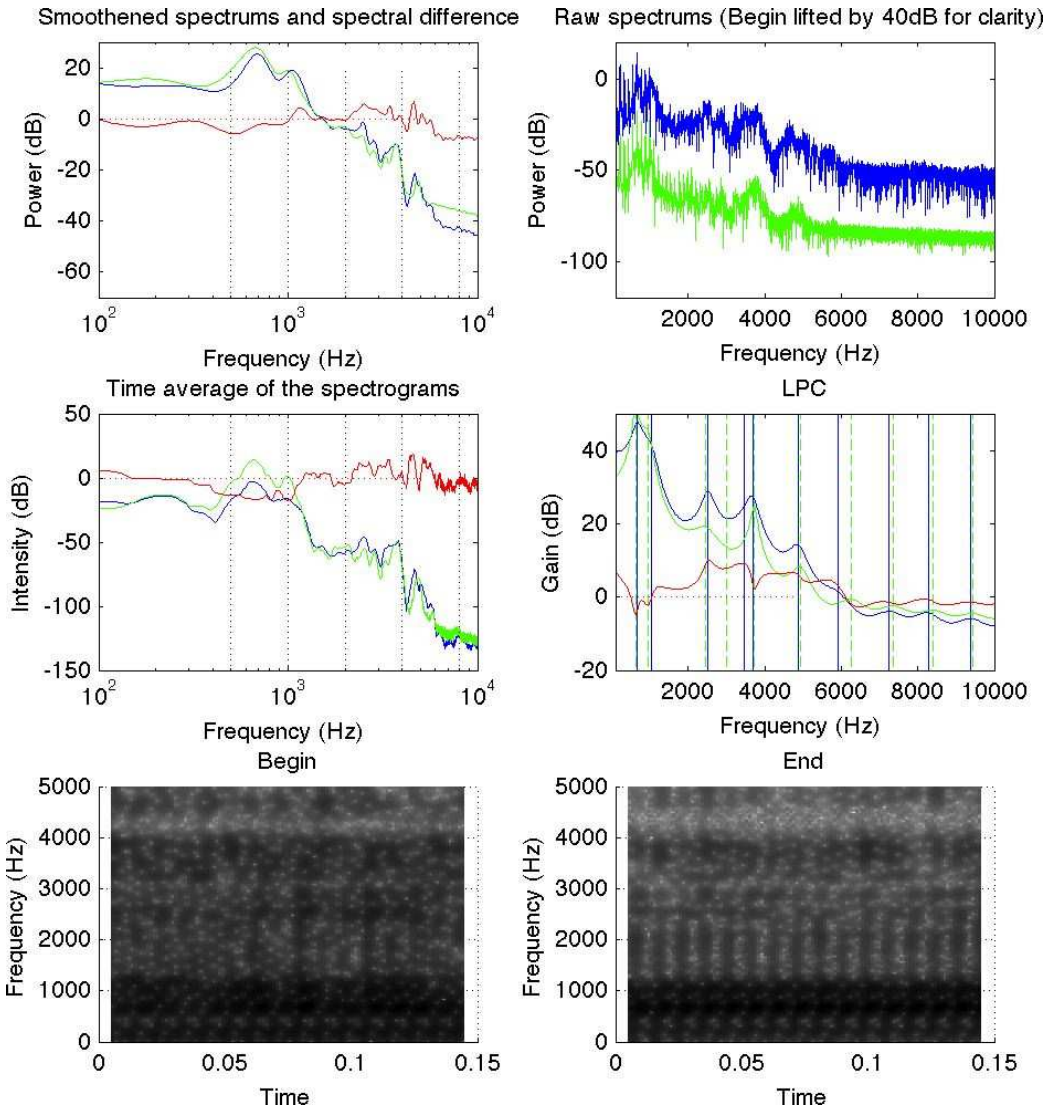


Figure 6.3: [ɑ] target f_0 =110 Hz. MRI sequence VIBE 1.8, duration 7.6 s.

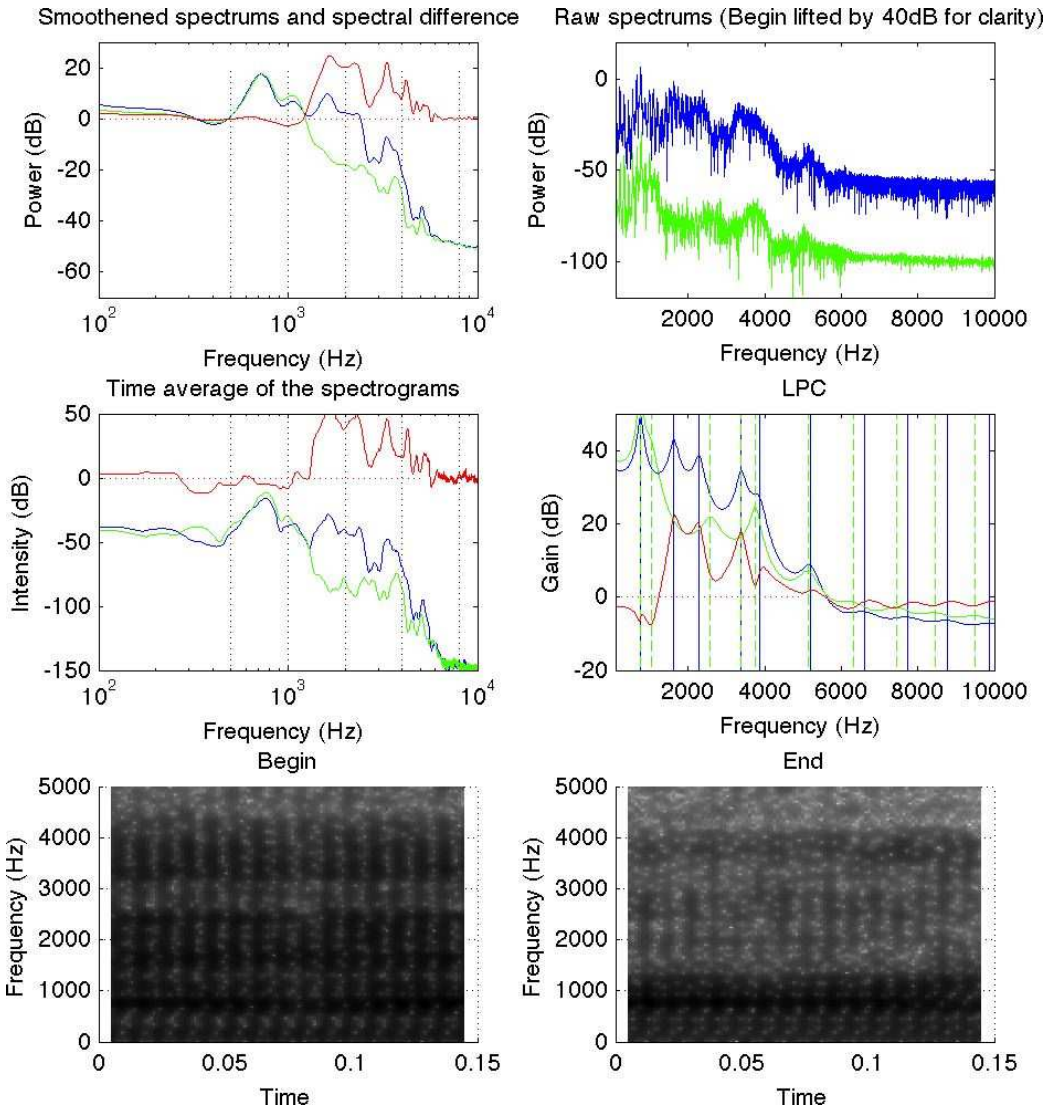


Figure 6.4: [æ-ɑ] target f_0 =110 Hz. MRI sequence VIBE 1.8, duration 7.6 s.

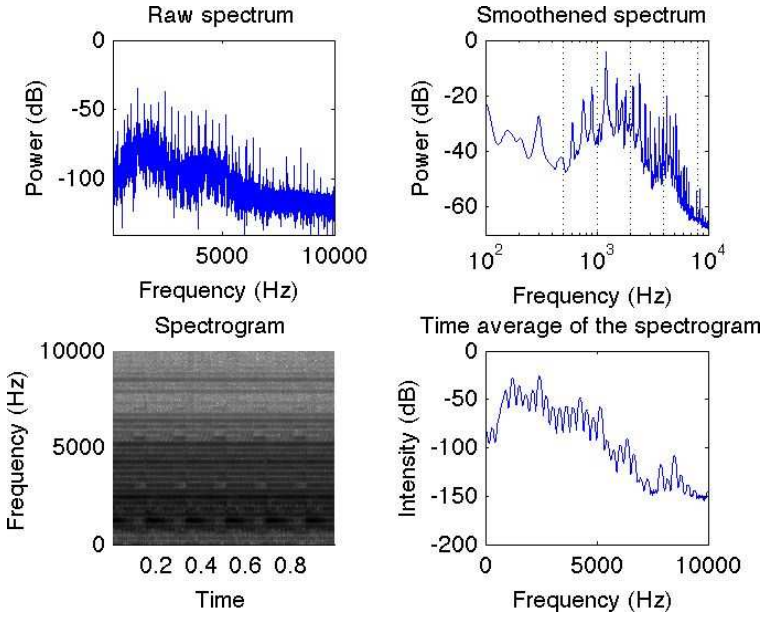


Figure 6.5: Acoustic noise charasteristics of a 8 s dynamic MR imaging sequence.

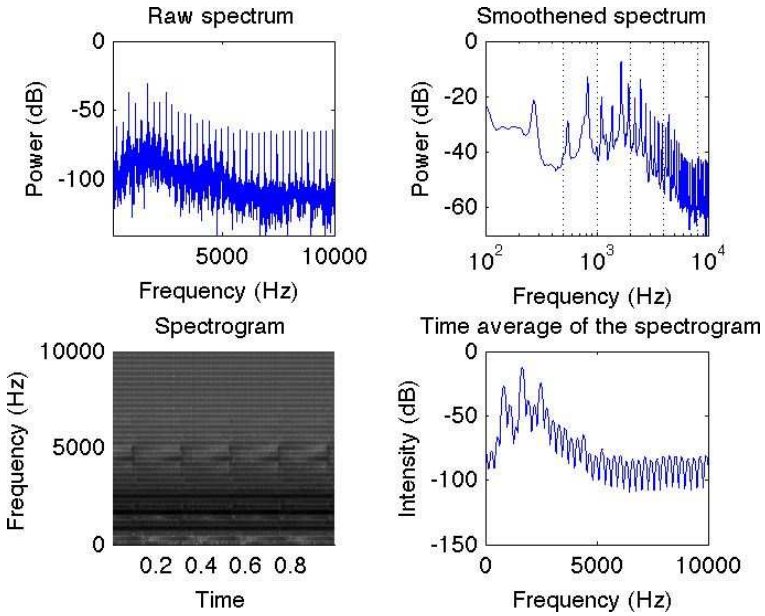


Figure 6.6: Acoustic noise charasteristics of the VIBE 1.8 MR imaging sequence

6.2 MRI data

The anatomical data obtained with MRI is of good quality. A typical example of a mid-sagittal cut from a 3D imaging series (VIBE 1.8) is shown in Figure 6.7. The data does not contain any significant artefacts and the air-tissue boundary is readily distinguishable in the images. This is true even in the nasal cavities most of the time as shown in Figures 6.7 and 6.8.

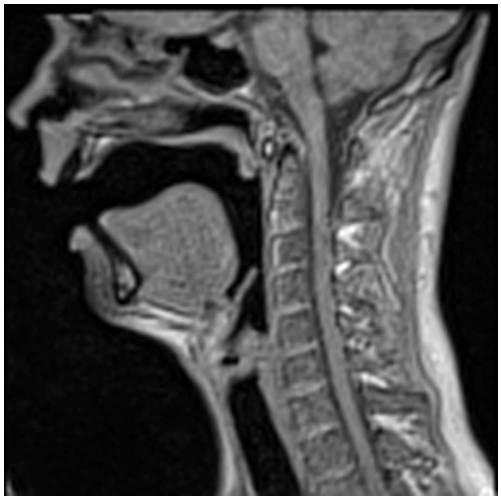


Figure 6.7: Sagittal section from a VIBE 1.8 image set of a production of the vowel [æ] with $f_0 = 110$ Hz.

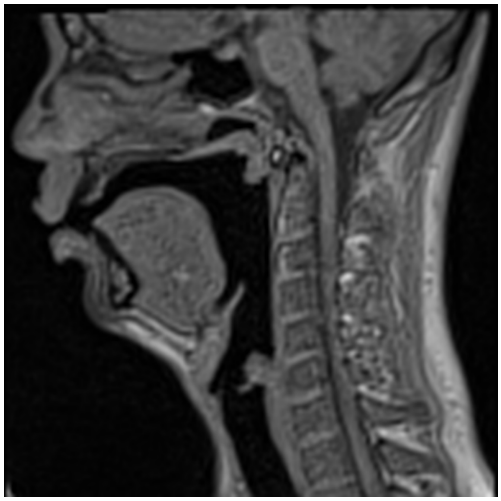


Figure 6.8: Sagittal section from a VIBE 1.8 image set of a production of the vowel [y] with $f_0 = 137.5$ Hz.

6.2.1 Data for developing a set of disqualification criteria

For data to be reliable it is necessary have rules when to disqualify a given sample. To develop a rejection criterium, we have conducted some comparisons to find out how changes in f_0 and formant frequencies are reflected in the anatomy and vice versa.

As an example of a clear disqualification case and of what a movement artefact would look like, we imaged a [æ-ɑ] glide with the static 3D sequence VIBE 1.8, see Figure 6.9. As expected an artefact is visible in the tongue contour. Similarly, the spectral difference is drastic as shown in Figure 6.4 (of the same production) and evident also in the formants listed in Table 6.2. The same glide was also imaged with a dynamic mid-sagittal MRI sequence. The first and final frame of the sequence are shown in Figure 6.11 which also shows an overlaid image illustrating the difference between the two frames.

Two further such comparisons were made: (1) The effect of changing f_0 from 110 Hz to 137.5 Hz; and (2) the effect of gravity on the articulators in a long production of a front vowel. These are shown in Figures 6.10 and 6.12 respectively. Figure 6.10 is based on two different VIBE 1.8 sets on vowel [ɑ] while Figure 6.12 is based on one dynamic sequence on [æ]. All of the combination images were produced with Photoshop CS5.

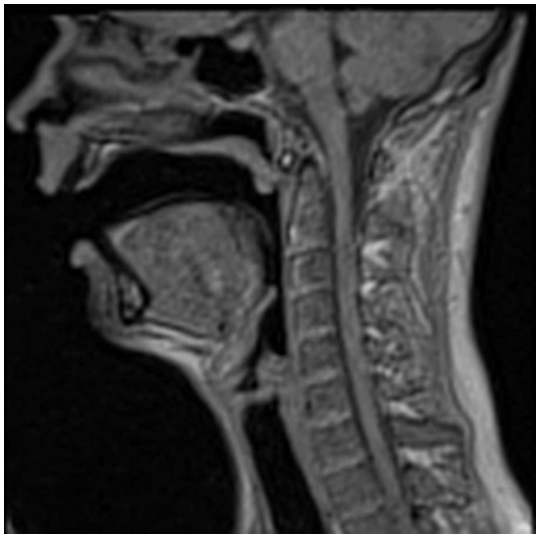


Figure 6.9: Static imaging of a glide: Sagittal section from a VIBE 1.8 image set of a production of the glide [æ-ɑ] with $f_0 = 110$ Hz.

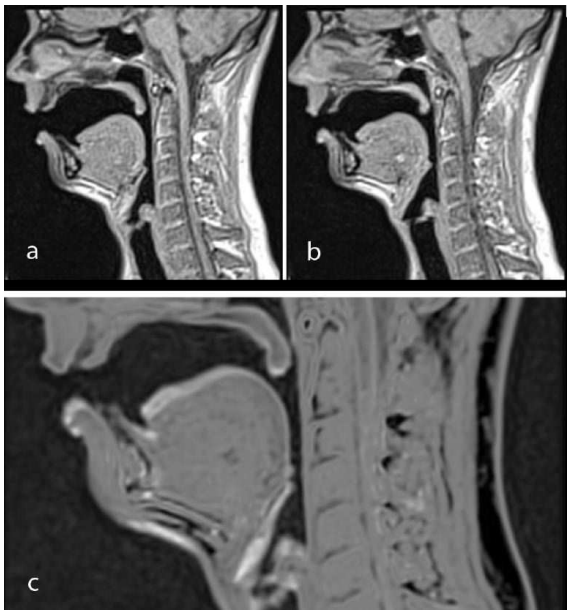


Figure 6.10: Sagittal sections of an 8 s production of the vowel [ɑ] with a) $f_0 = 110$ Hz; and b) $f_0 = 137.5$ Hz. An overlaid image is shown in c) to indicate their difference that is visible as lighter gray.

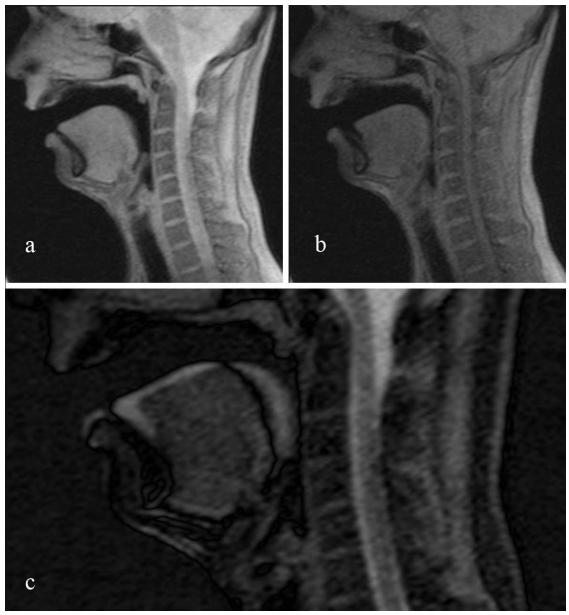


Figure 6.11: Sagittal sections from the a) beginning and b) end of an 8 s production of a [æ-ɑ] glide with $f_0 = 110$ Hz. An overlaid image is shown in c) to indicate their difference that is visible as lighter gray.

Figure 6.12: Sagittal sections from the a) beginning and b) end of an 8 s production of a [æ] with $f_0 = 110$ Hz. An overlaid image is shown in c) to indicate their difference that is visible as lighter gray.

6.3 Stability data on long vowel productions

Formant and f_0 extraction used the same algorithms and principles as that of the sound data from MRI (see Section 6.1 above). The variation in the samples was evaluated by computing the standard deviations (SD) of each of the acoustic measures with a moving 7.6 s window. Resulting SD data is illustrated in Appendix B. A visual inspection of the data and average curves shows that there is no consequential difference between the variances of the different groups.[1]

Based on this data optimal sampling times were sought for the whole sound data as well as the four partial sets: Short, long, $f_0 = 110$ Hz, and $f_0 = 137.5$ Hz data. The procedure continued by first averaging the standard deviation over each set. After this optimality was defined as the minimum of standard deviation in a given acoustic measure. A further combined optimum was calculated for the product of f_0 and formants F1-F4. The resulting optimal times for the data sets are listed in Table 6.1.

[1] A statistically significant difference may exist, but its size would be negligible.

What then is the shape of optimal sampling time distributions of the different measures? To find out, the optimums were calculated independently for each sample, and histograms were plotted of their time distributions. Figure 6.13 shows the distribution for the whole data. The histograms for the partial sets are located in Appendix B.

Table 6.1: Optimal beginning times (in seconds) for a 7.6 s sample.

Optimum on	**Data set**				
	Whole	Short	Long	110 Hz	137.5 Hz
Product	1.6	1.6	8.8	3.2	1.6
f_0	2.3	2.3	8.8	3.3	2.3
F1	1.6	1.4	1.9	2.4	1.4
F2	1.6	1.6	8.9	3.3	1.2
F3	2.3	2.3	1.7	3.3	1.6
F4	1.7	2.3	9.8	3.3	1.4

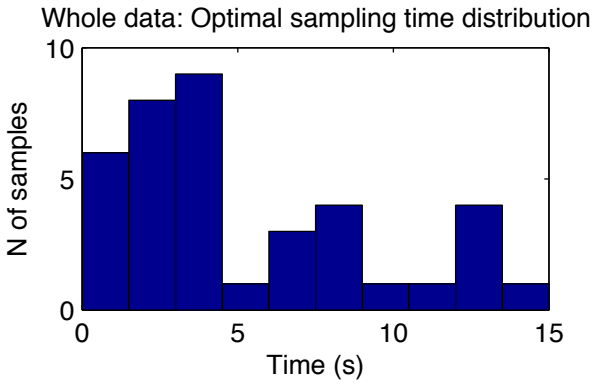


Figure 6.13: Optimal sampling time distribution for the whole data.

Table 6.2: .

]f_0s, formants F1-F4, formant distances and sound pressure level differences for vowel productions with target $f_0 = 110$ Hz. Glides with [1] Vibe 1.8 and [2] dynamic imaging sequences. Aliased F2 values are shown in [square brackets].

Sound		[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]	[æ-ɑ][1]	[æ-ɑ][2]
f_0 (Hz)	begin	109.2	108.1	109.4	106.8	109.4	108.4	106.3	107.6	108.4	109.2
	end	109.7	108.6	109.7	108.9	113.1	110.5	107.6	110.2	109.4	110.5
F1 (Hz)	begin	669	471	244	408	293	271	783	416	756	773
	end	631	498	252	420	285	296	776	448	742	753
F2 (Hz)	begin	1041	1891	[956]	739	669	1585	1621	1515	1626	1510
	end	964	1929	[724]	704	661	1533	1303	1389	1051	1037
F3 (Hz)	begin	2508	2605	2438	2283	2178	2053	2292	2013	2285	2312
	end	2460	2356	2074	2168	2151	1999	2375	2088	2578	2447
F4 (Hz)	begin	3464	3329	3223	3018	3316	3239	3075	2964	3386	3494
	end	3000	3140	3007	2875	3011	3054	3474	3206	3377	3418
Formant distance (Hz)		280	316	484	187	266	201	517	285	645	498
SPL difference (dB)		5.9	8.3	0.5	6.4	3.5	4.3	4.1	2.2	-0.2	0.5

Table 6.3: .

]f_0s, formants F1-F4, formant distances and sound pressure level differences for vowel productions with target $f_0 = 137.5$ Hz. Aliased F2 values are shown in [square brackets].

Sound		[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
f_0 (Hz)	begin	137.0	138.2	136.1	137.0	135.3	130.9	134.9	136.1
	end	135.3	139.6	136.5	138.2	137.0	135.3	137.4	137.0
F1 (Hz)	begin	608	491	286	552	371	324	739	446
	end	634	570	320	558	374	293	743	508
F2 (Hz)	begin	1015	[747]	[1025]	869	549	1514	1501	1391
	end	994	1934	[1060]	846	807	1468	1202	1243
F3 (Hz)	begin	2394	1931	2246	2309	2016	2005	2216	2009
	end	2268	2426	2037	2310	2065	1996	1489	2052
F4 (Hz)	begin	3435	2612	3174	2810	3317	3180	2342	3153
	end	3406	3274	2875	3154	2970	3211	2364	3196
Formant distance (Hz)		134	268	368	344	337	64	151	172
SPL difference (dB)		3.2	8.6	1.1	3.6	5.0	-0.1	4.8	2.4

Chapter 7

Discussion

We have described mathematical and numerical modeling of vowel production as well as experimental protocols, MRI sequences, and a sound recording system that can be used for simultaneous sound and anatomical data acquisition of human speech. The results and experiences of a pilot experiment on vowel formants and the quality of the corresponding anatomical data have been reported. Such data sets are intended for parameter estimation, fine tuning, and validation of a mathematical model for speech production. However, these methods and technologies have a wide range of other applications in phonetics and medicine as well. It is now time to make observations on the results and give recommendations based on the experience gained. Some of the findings in this Chapter have been published in our articles cited previously in this thesis.

7.1 Observations on the results

7.1.1 Modeling

The computed formants F1 to F4 in Chapter 3 differ from the corresponding measured formants by $3\frac{1}{2}$ semitones. Having said that, the *ratios* between the computed formants and the measured formants match each other very well.

There is a simple physical explanation why such a discrepancy is to be expected. In Equations (2.2), we use the Dirichlet boundary condition on the lip opening. This results in a vibrational node at the opening. In reality, such a node would appear further away outside the mouth in a frequency dependent position. In that sense, the real life VT is effectively longer than the one described by Equations (2.2), resulting in lower formants. While it

is clear that the surrounding acoustic space does affect the resonance structure of the VT, without results from modeling and measurements the exact nature of the lengthening effect remains subject to hypothesis. Therefore, we should also model the surrounding acoustic space or the corresponding radiation load.

7.1.2 Recording setup

Our recording system does not cause data samples to be disqualified and all acoustic artefacts are readily identifiable – by additional measurements and modelling where required. Its main shortcoming at the moment is the 1 kHz peak discussed below.

In constructing and measuring the setup, we use sound sources which are not completely linear. However, their nonlinear behaviour is restricted to certain exceptional frequencies – most notably the resonances of the point-wise sound source. There are not many of these frequencies and they have been listed.

There are comparable efforts for recording speech during MRI. For example, Ericsdotter (2005) used an optical differential microphone system. In comparison with the optical system, the acoustic system presented in this work is larger, and acoustic impedances need more attention. Yet, the acoustic equipment is always linear if used within the operational limits of the electronic parts, the sound collector is immune to structural vibrations, and the system is readily modifiable.

7.1.3 Sound data from MRI

The subject was instructed to keep f_0 in a given reference value, and he was able to do that with an error of $\pm$ 3.5 Hz in all experiments. The sound pressure given by the subject almost always increases towards the end of the sample.

It is hard to find a pattern in 6.1 except in one case: The [æ] productions are clearly similar to those of the [æ-ɑ] glides. This makes it necessary to be cautious about the quality of the corresponding static vowel data.

The formant extraction with LPC does not always work as well as one would hope. Examples can be seen in Figure 6.1 – as well as the corresponding data in Tables 6.2 and 6.3. The [e] with f_0 = 137.5 Hz clearly has an aliased F2 in the begin sample and both [i]s have an aliased F2 in begin and end samples.

This may be due to the fact that there appears a somewhat systematic

peak at about 1 kHz in the spectrums of the MRI sound data (see the Figures in Appendix A). This peak is not dependent on f_0 nor on the vowel in question. Thus it has to be concluded that it most likely is an artefact of the system. At the time of writing, its cause is still unclear. The most likely cause for the artefact are reflections from the surfaces inside the MRI registration coil. Careful measurements of the system – going through the whole calibration sequence – will have to be conducted to clarify the situation.

The formant distance measure is not as robust as a disqualification criterion should be. In particular, it does give a low value, 64, for [y] with f_0 = 137.5 Hz, see Table 6.3. But at the same time the measure gives a comparably low value, 151, to [æ] with f_0 = 137.5 Hz, which can be seen in Figure 6.1 to be about as shifted as the intentional glides. Thus the measure gives good (low) values to both shifted and unshifted productions.

As can be expected, the sound energy in the MRI noise is concentrated on few discrete frequencies, see Figures 6.5 and 6.6. As the imaging noise is uncorrelated with the speech, it can be removed from a speech sample by subtraction in the spectral domain before formant extraction.

7.1.4 Anatomical data

The MR images have generally a very good air-tissue contrast with mostly continuous air-tissue contours. However, this does not mean that all continuous contours should be considered reliable representations of the anatomy. The reason for this is that the VIBE-sequence is provided as a black box without information on how it builds the image. How the image is built in a particular sequence determines if and how a given type of movement is visible in the images – the worst case being that an image does not show any kind of artefact even though the subject moved significantly during the sequence.

We intentionally produced a motion artefact as can be seen in Figure 6.9. However, motion does not always produce an artefact in the images – at least not in the form of a discontinuous air-tissue contour. This can be seen in the continous but somewhat unnatural tongue contour in Figure 6.7. This contour is most likely the product of the shift which is made evident by the shift in that productions F2 value, see Figure 6.1.

The anatomical data presents a complex pattern of movements (see Figures in Section 6.2) which are probably caused by several different reasons. One obvious reason is the effect gravity has on the position of articulators when the subject is in a supine position. This effect has been studied

and reported in detail previously by Engwall (2006) and Stone et al. (2007).

Another reason is the subject's lungs slowly empty during a long production. This results in lowering of the thorax which in turn affects the position of the articulators – in particular, that of the the larynx and the vocal folds (Yanagihara and Koike 1967). Furthermore, control of the subglottal pressure in long productions is a complex issue. Aiming to produce a constant f_0 may cause the subject to move his larynx (Stevens 1998).

7.1.5 Stability of long vowel productions

The deviation data on vowel productions shows a clear result in two ways: First, the average deviation curves plotted in Appendix B show that the standard deviation of a 7.6 s sample stabilises to a fairly constant level in about a second from the beginning of the production.

Second, the histogram on the whole data in Figure 6.13 shows the highest frequencies in classes under 5 s. However, this is only due to the short data being limited by their length to sample beginning times of under 5 s. Thus it emerges that there is no particular pattern to the optimal sampling time distributions expect that of a uniform random distribution.

7.2 Recommendations and future directions

7.2.1 Model validation

Most of the work that has gone to producing this thesis has been concerned with making validation of the mathematical and numerical models of vowel production possible. Let us now look at the questions of what to validate and in what way.

The validation of the mathematical model can be done only indirectly by validating a simulator based on it. There are at least two simulators – on already in existence, the other yet to be written – that can be validated: First, we can validate the resonance model. This means comparing formants measured from sound samples to the resonances produced by the resonance model described in Chapter 3. Second, we can validate a time dependent vowel model. This means comparing formants measured from sound samples to formants measured synthesised vowel sounds produced by the time dependent vowel model.

The issue of how to validate, involves choosing the measures for accuracy of the simulation. Here we have two relevant contexts: That of a human listening to the speech and that of simple physical accuracy of

the simulation. If we opt for validation with the human hearing system in mind as the ultimate measure of the success of the system, then the measures have to be derived from the properties of that system. Another approach would be to use the simplest measures that we can find, that give a reasonable guarantee of consistency of the simulation.

7.2.2 Data acquisition

For successful data acquisition, it is necessary to identify relevant parameters concerning the whole experimental setting that must be kept constant and well-documented at all times. It is not always a priori clear what should be regarded as a relevant parameter, or what practical steps must be taken to keep them under control.

Now, the relevant parameters can be divided into two groups: (1) *physiological parameters* involving the human subject; and (2) *physical parameters* of the measurement equipment. We have worked extensively to standardize the latter group as reported in Chapter 5. On the other hand, the first group of parameters still has several open questions.

Figure 6.10 indicates that different levels of f_0 result in visible differences of vocal tract configuration while uttering [ɑ]. Changes in sound pressure may result in a similar change as in Figure 6.10. To exclude this, the subject tried to keep the sound pressure same in both imaging sessions, but he did not receive any feedback in that respect. However, the (subjective) exhalation time was of the same length in both measurements.

It should be noted that the formants are considered as purely acoustic parameters of the vocal tract geometry, and — as such — they do not depend directly on dynamic variables such as the sound pressure and the air flow but, instead, through detectable changes in vocal tract geometry. It is not clear whether one should (for physically motivated reasons) aim at constant sound pressure or at constant air flow in a measurement leading to Figure 6.10. However, observing the flow inside the MRI machine is probably very challenging.

The phonetic and articulatory problems stem from the inability of a subject to maintain a stable vocal tract shape for the duration of the MRI sequnce. That is, there is a trade-off between image quality in terms of resolution and speech production quality in terms of articulatory stability. The duration of the 3D MRI scan we used is 7.6 s, and the sound recording time is ≈ 2 s longer than that; see Figure 4.2. The subject should be able to maintain constant position, configuration of the vocal organs, all sound characteristics, and the type of phonation during the whole period. Ac-

cording to our experience, this is a difficult requirement even for a healthy subject.

Our work indicates that the problems cannot be circumvented altogether. There are, however, simple means to further improve articulatory stability during recordings. We propose, at least, the following:

1. The test subject should be familiar with the MRI noise as well as the cue signal so as to perform optimally in the experimental situation.
2. The intensity of the cue signal should match the MRI noise so that the initial voice production can be maintained, i.e., possible sound intensity fluctuation should be avoided. Since the earphones of the MRI machine cannot produce enough volume on low frequencies to match the desired level, the cue has to contain energy on high spectrum as well without changing the perceived f_0.
3. The MRI noise itself is periodic and interferes with the perception voice f_0 when they are close. Hence, the cue should avoid the peaks in the MRI noise frequency profile.
4. f_0 should be standardized separately for each test subject taking into account their comfort zone.
5. The cue signal should be longer to allow the subject more time to inhale.
6. The time interval between speech and MRI onsets should be made longer. This, however, lengthens the required vowel duration.
7. A single imaging session should not last longer than 0.5 hours. If more imaging time is needed on a single subject breaks have to be introduced.

7.2.3 Other observations

Finally, there are some points to take into consideration concerning the experimental setting. First, the acoustic MRI noise is significantly different for different imaging sequences. Ultrafast sequences — such as 3D VIBE used in this work — require maximal performance of gradient system both in terms of slew rate and amplitude. This results in exceptionally loud acoustic noise. We remark that even smallest changes in parameters of a given MRI sequence may change acoustic noise significantly. It is thereby essential to maintain sequence parameters and patient positioning constant.

An obvious short coming of our anatomical data is that at the moment it does not incorporate the teeth in any form. We have conducted an experimental acquisition based on the work reported in Takemoto et al.

(2004). The results, however, were not satisfactory. In particular, this method seems usable for studies involving a limited number of subjects as it does require a good deal of manual fine tuning. An approach which would be more easily automated would be using MRI-visible markers on the teeth and acquiring the teeth model by e.g. traditional casts or modern digital scans.

As for the sound data, a good disqualification measure remains an open question. The LPC based formant distance measure is a good starting point, but it does need a more reliable formant extraction algorithm to work properly. One way to improve the extraction would be to use the a priori knowledge we have on the speech sound that the subject is producing. Namely, the MRI data on the same production. The probable formant frequencies can be estimated by extracting the vocal tract geometry from the MRI data and running the geometry through the VT resonance simulator. These frequency estimates can then be used to prime formant extraction from the sound data, thus avoiding obviously false formant values and giving a more exact measure for any possible drift in the geometry.

A necessary way to improve the reliability of our data is to monitor the sound of the subject throughout the whole production and not just at the begining and end. This requires implementing DSP filtering of the samples to remove the noise of the MR imaging sequence.

7.3 Conclusion

All in all, the results in this thesis are encouraging. They clearly point to directions where the setup refinements and better understanding will iteratively approach a useful solution to the linked problem of modelling and measurement.

No actual production of a speech sound can be infinite in length since lung volume is finite. The only way to hold the articulators stationary for long periods of time is not to actually produce a sound while doing so – to just imagine producing the sound and trying to hold the anatomical configuration stable. But if the subject does not produce any sound, how can we know if he or she is actually holding the articulators in the desired position? Thus it is necessary to find a good trade of between sound production length and stability. Only in this way is it possible to acquire a matching sound-anatomy pair and facilitate the validation of any production models.

It is true that it is impossible to be certain that a given production was stable enough. The important thing is to recognise error sources and the

ways in which different types of errors can be identified. And make every effort to minimise all errors.

Bibliography and References

Aalto, A. (2009). A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load. Master's thesis, TKK, Helsinki. Available at `http://math.tkk.fi/research/sysnum/`.

Aalto, A., Alku, P., and Malinen, J. (2009). A LF-pulse from a simple glottal flow model. In *MAVEBA 2009*, pages 199–202, Florence, Italy.

Aalto, D., Malinen, J., Palo, P., Aaltonen, O., Vainio, M., Happonen, R.-P., Parkkola, R., and Saunavaara, J. (2011a). Recording speech sound and articulation in MRI. In *Biodevices 2011*, pages 168 – 173, Rome, Italy.

Aalto, D., Malinen, J., Palo, P., Saunavaara, J., and Vainio, M. (2011b). Estimates for the measurement and articulatory error in MRI data from sustained phonation. To appear in *Proceedings of ICPhS 2011*, Hong Kong, China.

Alku, P., Horáček, J., Airas, M., Griffond-Boitier, F., and Laukkanen, A.-M. (2006). Performance of glottal filtering as tested by aeroelastic modelling of phonation and FE modelling of vocal tract. *Acta Acustica united with Acustica*, 92:717–724.

Atal, B. S. and Schroeder, M. R. (1967). Predictive coding of speech signals. In *Proceedings of IEEE Conference on Communication and Processing*, pages 360 – 361.

Blackburn, C. S. (1996). *Articulatory Methods for Speech Production and Recognition*. PhD thesis, Trinity College Cambridge & Cambridge University Engineering Department.

Chiba, T. and Kajiyama, M. (1941). *The Vowel, Its Nature and Structure*. Phonetic Society of Japan.

Deufhard, P., Weiser, M., and Zachow, S. (2006). Mathematics in facial surgery. *Notes of the AMS*.

Engwall, O. (2000). Are statical mri data representative of dynamic speech? results from a comparative study using mri, ema and epg. In *Proceedings of International Conference on Spoken Language Processing 2000 (ICSLP 2000)*, pages I: 17–20.

Engwall, O. (2003). A revisit to the application of mri to the analysis of speech production - testing our assumptions. In *Proceedings of the 6th Int Seminar on Speech Production*.

Engwall, O. (2006). *Speech production: Models, Phonetic Processes and Techniques*, chapter Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation., pages 301 – 314. Psychology Press, New York.

Engwall, O. and Badin, P. (1999). Collecting and analysing two- and three-dimensional MRI data for swedish. *TMH-QPSR*, (3-4/1999):11–38.

Ericsdotter, C. (2005). *Articulatory-Acoustic Relationships in Swedish Vowel Sounds*. PhD thesis, Stockholm University, Stockholm, Sweden.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Fant, G., Liljencrants, J., and Lin, Q. (1986). A four-parameter model of glottal flow. Technical report, QPRS: Dept. for Speech, Music and Hearing, Stockholm.

Fetter, A. and Walecka, J. (1980). *Theoretical Mechanics of Particles and Continua*. McGraw–Hill, New York.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception*. Springer-Verlag.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47:1202.

Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (2006). Formants and vowel sounds by finite element method. In *The Phonetics Symposium 2006*, pages 24 – 33, Helsinki, Finland.

Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (2007). Vowel formants from the wave equation. *Journal of the Acoustical Society of America Express Letters*, 122(1):EL1–EL7.

Helmholtz, H. L. F. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Braunschweig: F. Vieweg.

Ishizaka, K. and Flanagan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal*, 51:1233–1268.

Johnson, C. (1987). *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press.

Lukkari, T. and Malinen, J. (2011a). Webster's equation as a conservative linear system. *Manuscript*.

Lukkari, T. and Malinen, J. (2011b). Webster's model with curvature and dissipation. *Manuscript*.

Lukkari, T., Malinen, J., and Palo, P. (2007). Recording speech during magnetic resonance imaging. In *MAVEBA 2007*, pages 163 – 166, Florence, Italy.

Lukkari, T., Malinen, J., and Palo, P. (2008). Puheen äänittäminen magneettiresonanssikuvauksen aikana. In *Fonetiikan päivät 2008*, pages 57 – 64, Tampere, Finland.

Malinen, J. and Palo, P. (2009). Recording speech during MRI: Part II. In *MAVEBA 2009*, pages 211–214, Florence, Italy.

Malinen, J. and Staffans, O. J. (2006). Conservative boundary control systems. *J. Diff. Eq.*, 231(1):290 – 312.

Moisik, S. (2008). A three-dimensional model of the larynx and the laryngeal constrictor mechanism: Visually synthesizing pharyngeal and epiglottal articulations observed in laryngoscopy. Master's thesis, University of Victoria, Victoria, Canada.

Rofsky, N., Lee, V., Laub, G., Pollack, M., Krinsky, G., Thomasson, D., Ambrosino, M., and Weinreb, J. (1999). Abdominal MR imaging with a volumetric interpolated breath-hold examination. *Radiology*, 212(3):876 – 884.

Saad, Y. (1992). *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, Manchester.

Sondhi, M. M. (1986). Resonances of a bent vocal tract. *Journal of the Acoustical Society of America*, 79(4):1113–1116.

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press.

Stone, M., Stock, G., Bunin, K., Kumar, K., Epstein, M., Kambhamettu, C., Li, M., Parthasarathy, V., and Prince, J. (2007). Comparison of speech production in upright and supine position. *Journal of the Acoustical Society of America*, 122(1):532 – 541.

Takemoto, H., Kitamura, T., Nishimoto, H., and Honda, K. (2004). A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions. *Acoustic Science and Technology*, (25):468–474.

Titze, I. (2008). Nonlinear source-filter coupling in phonation: Theory. *Journal of the Acoustical Society of America*, 123(5):2733–2749.

Švancara, P. and Horáček, J. (2006). Numerical modelling of effect of tonsillectomy on production of Czech vowels. *Acta Acustica united with Acustica*, 92:681 – 688.

Yanagihara, N. and Koike, Y. (1967). The regulation of sustained phonation. *Folia phoniatrica*, 19:1 – 18.

Appendix A

Sound data

On the following pages are the spectrum and spectrogram Figures for the MRI samples of the Finnish vowels [ɑ, e, i, o, u, y, æ, ø] with $f_0 = 110$ and 137.5 Hz and [æ-ɑ] glides with $f_0 = 110$ Hz. The color code for the curves is as follows: Blue corresponds to the begin sample, green to the end sample and red to their spectral difference. In the LPC plots the solid blue vertical lines represent the estimated formants of the begin sample and the dashed green vertical lines those of the end sample. In the spectrum plots, dotted vertical lines simply mark the frequencies 500, 1000, 2000, 4000, and 8000 Hz while the dotted horizontal line marks the 0 dB level.

Also listed under each Figure are the following acoustic measures: Sound pressure level difference between the begin and end samples in dB, estimated fundamental frequency f_0 in the begin and end samples, and formant distance between the samples. A detailed description of the algorithms used in generating the Figures and calculating the measures can be found in Chapter 6.

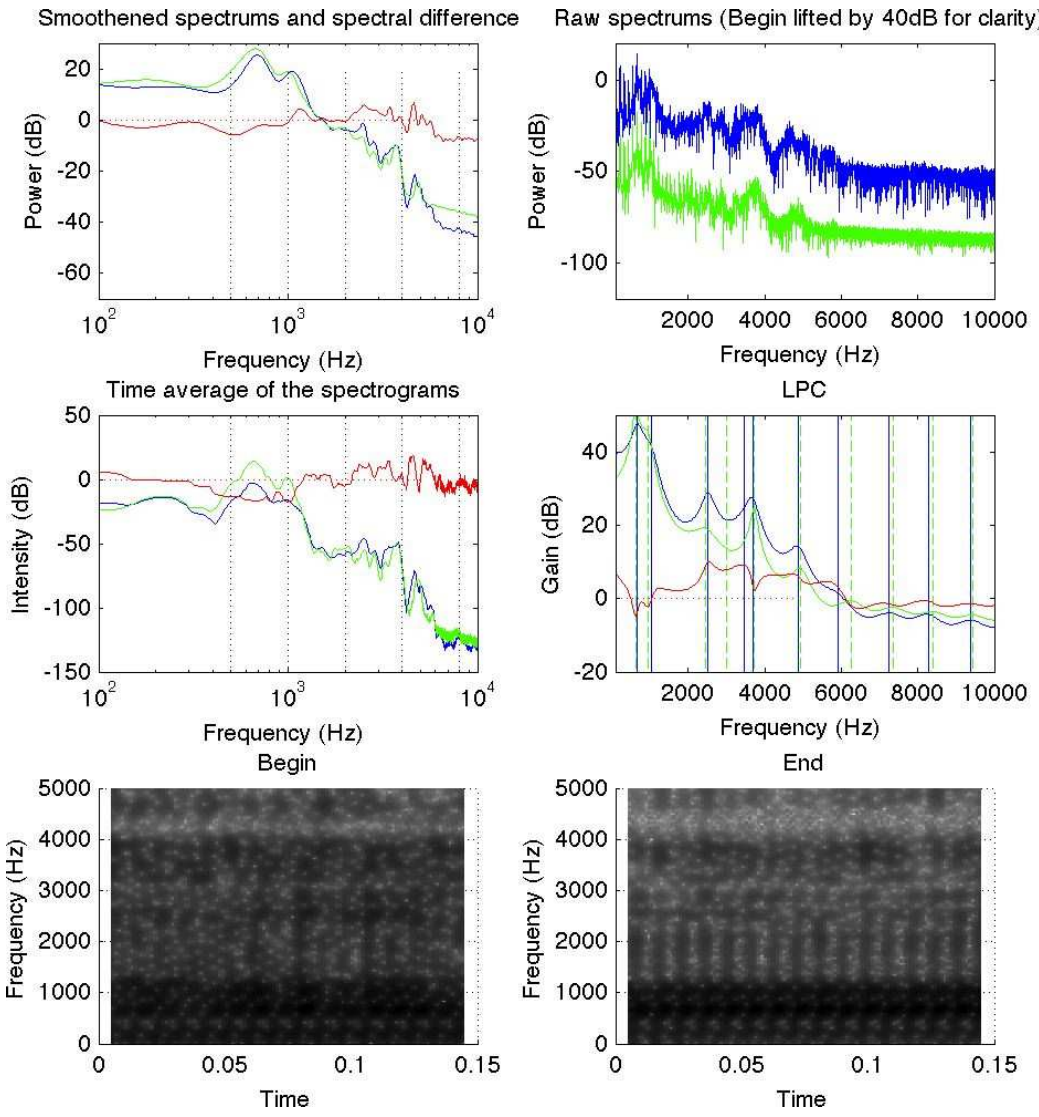


Figure A.1: [ɑ] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6s.

Sound pressure level difference: 5.9
Fundamental frequency in the begin sample: 109.2
Fundamental frequency in the end sample: 109.7
Formant distance: 280

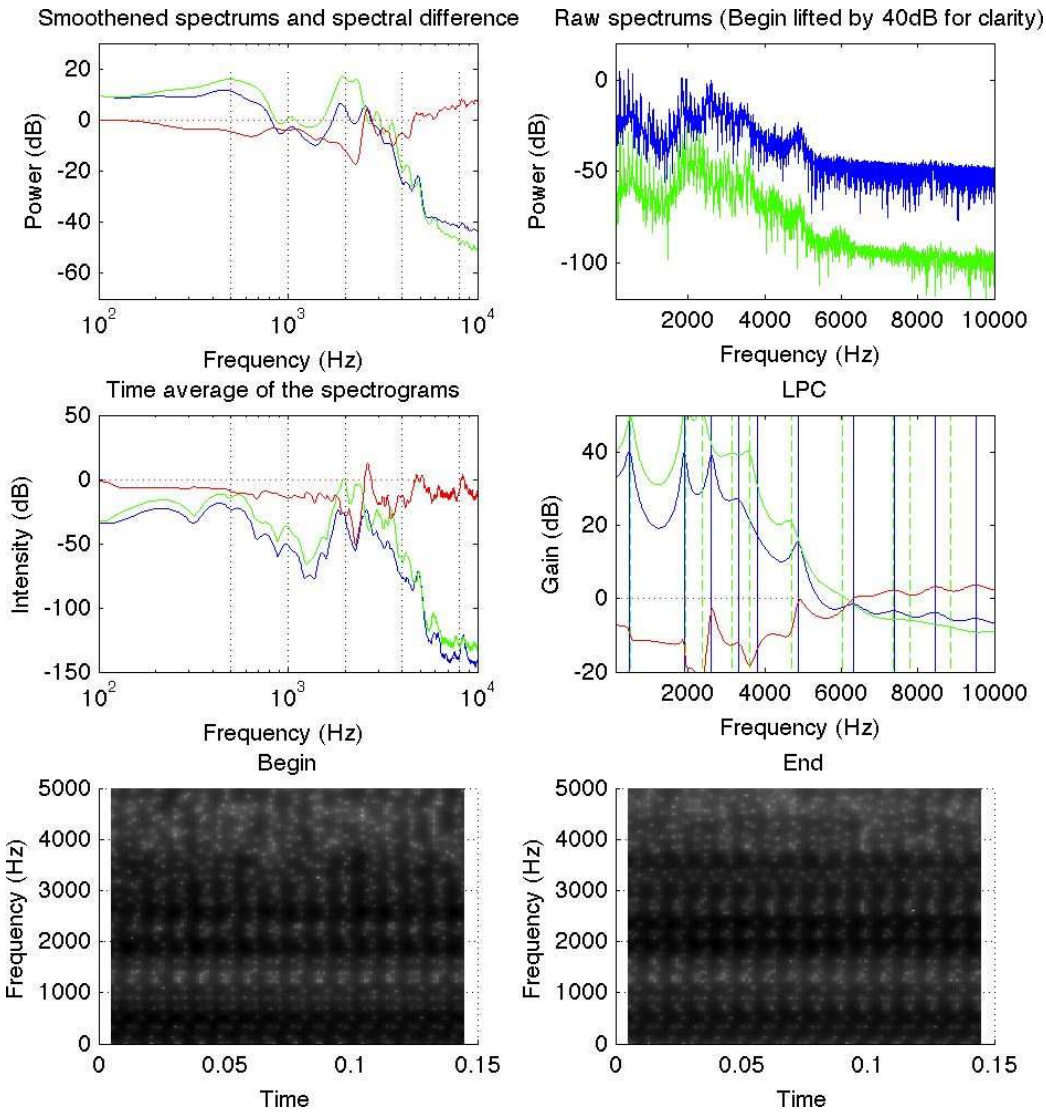


Figure A.2: [e] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 8.3
Fundamental frequency in the begin sample: 108.1
Fundamental frequency in the end sample: 108.6
Formant distance: 316

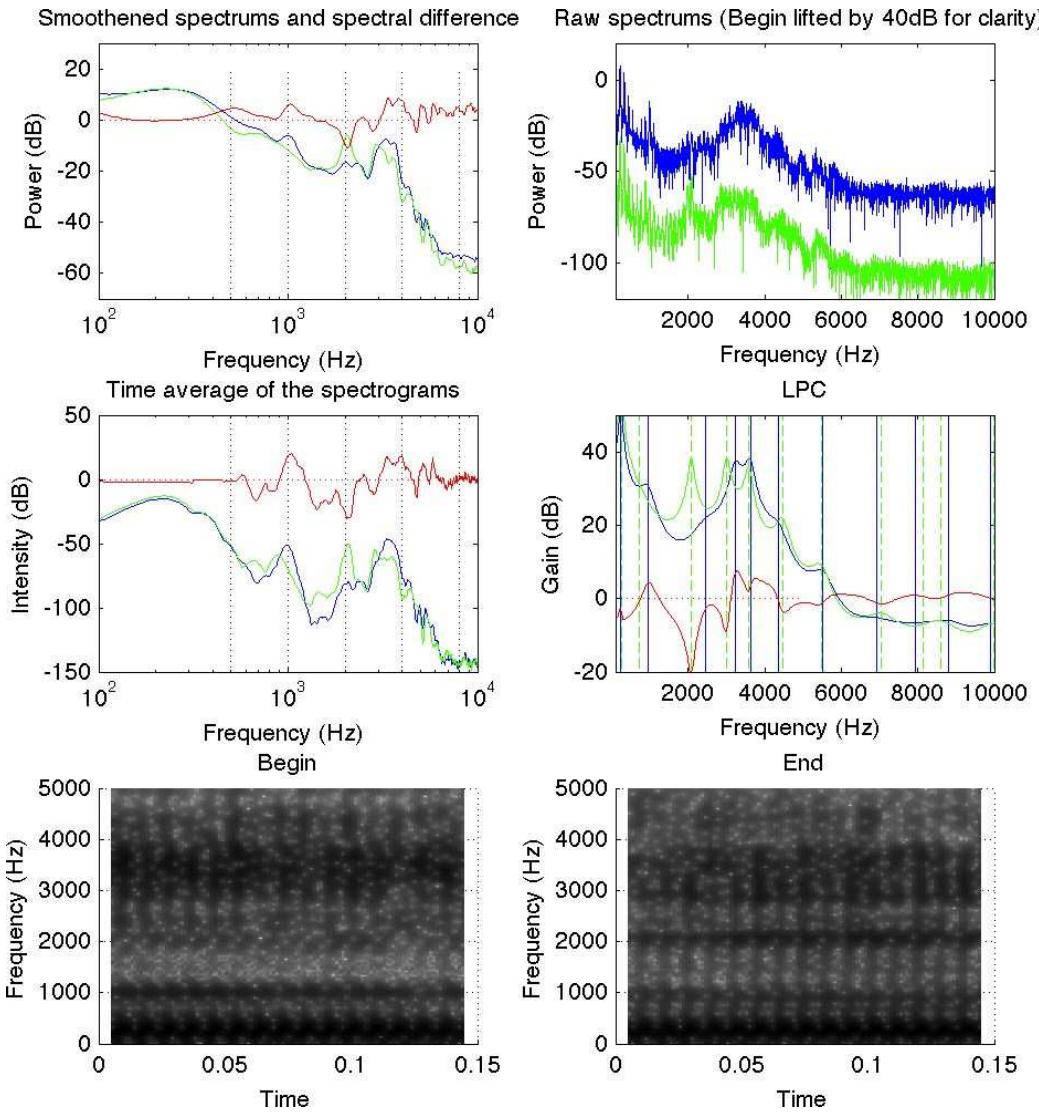


Figure A.3: [i] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 0.5
Fundamental frequency in the begin sample: 109.4
Fundamental frequency in the end sample: 109.7
Formant distance: 484

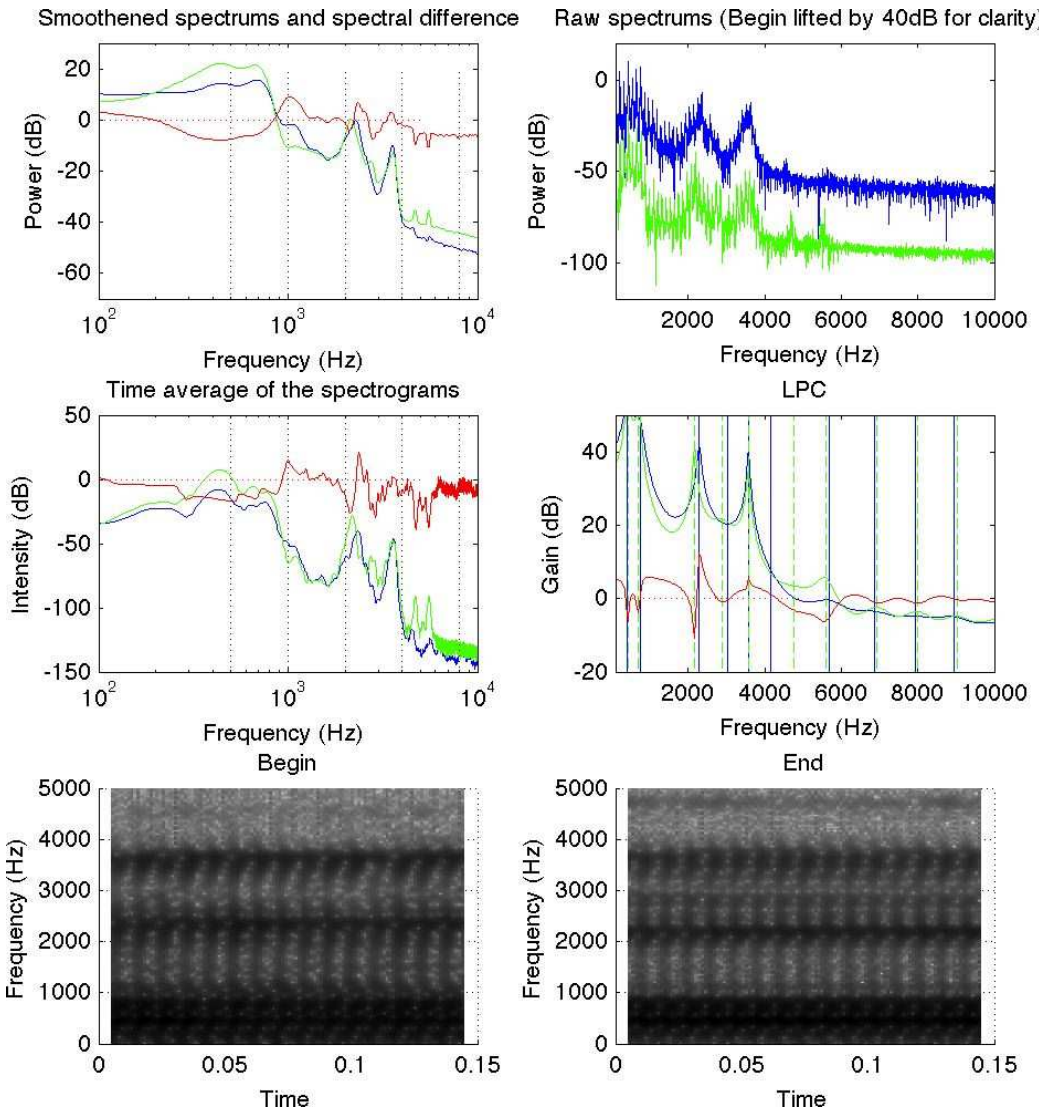


Figure A.4: [o] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 6.4
Fundamental frequency in the begin sample: 106.8
Fundamental frequency in the end sample: 108.9
Formant distance: 187

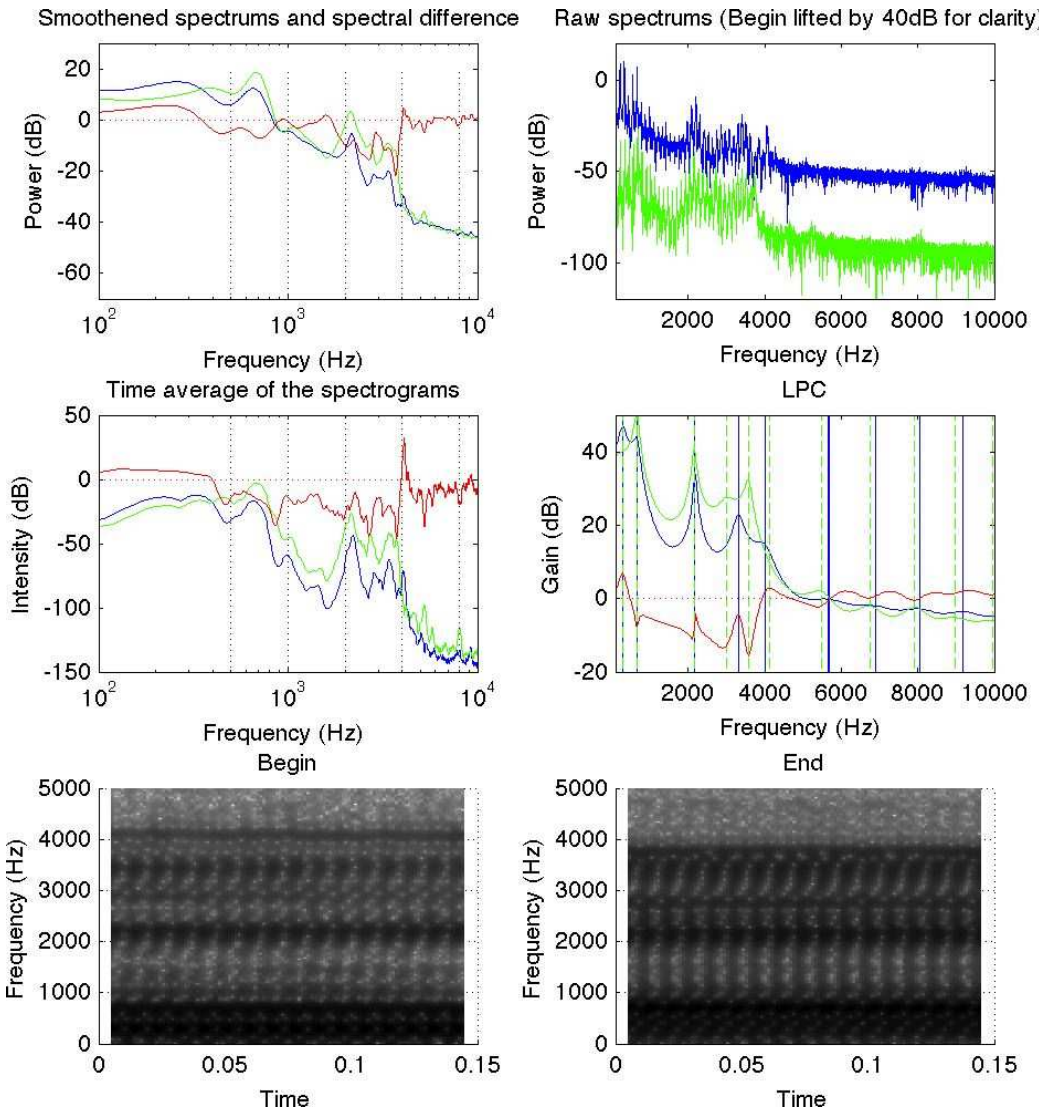


Figure A.5: [u] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 3.5
Fundamental frequency in the begin sample: 109.4
Fundamental frequency in the end sample: 113.1
Formant distance: 266

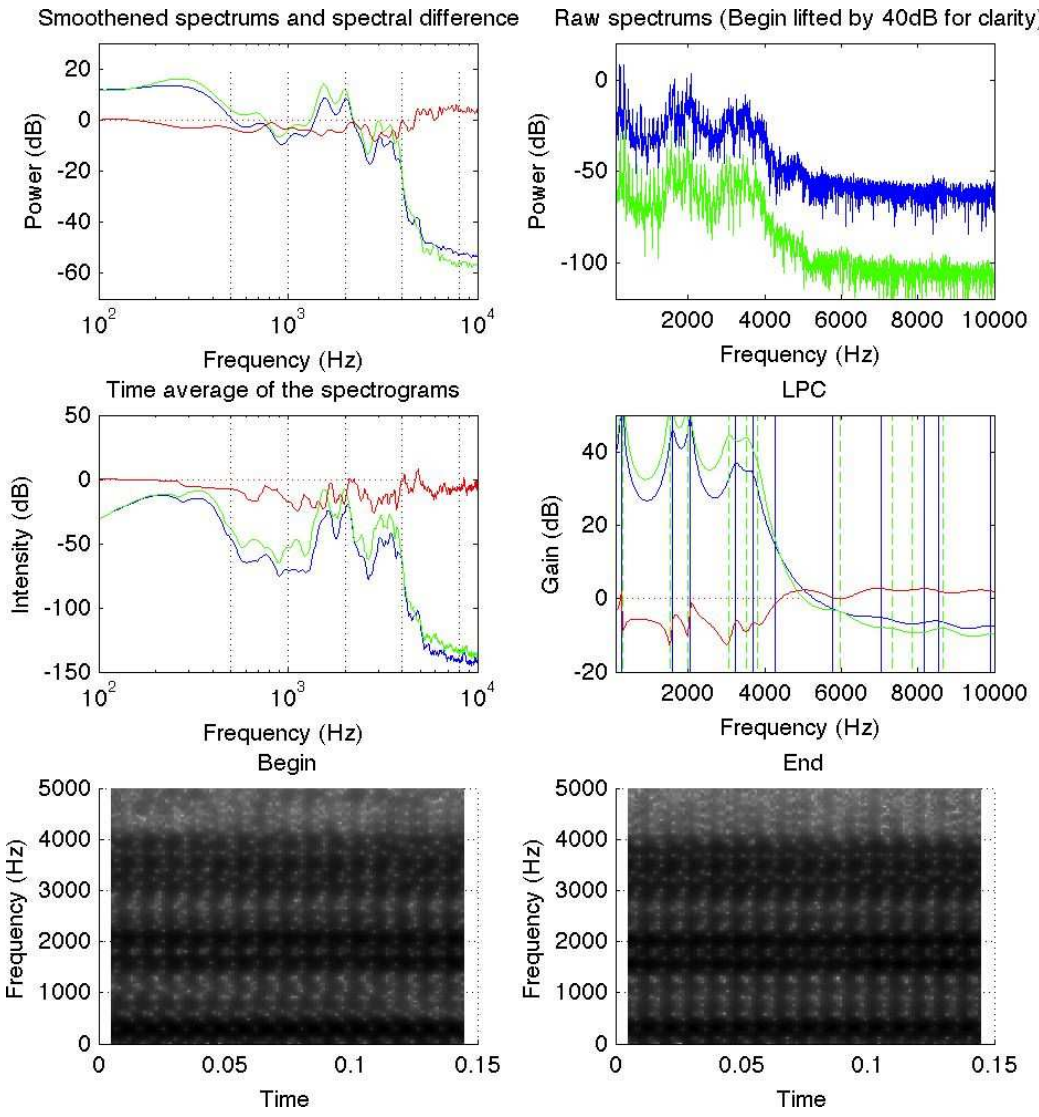


Figure A.6: [y] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 4.3
Fundamental frequency in the begin sample: 108.4
Fundamental frequency in the end sample: 110.5
Formant distance: 201

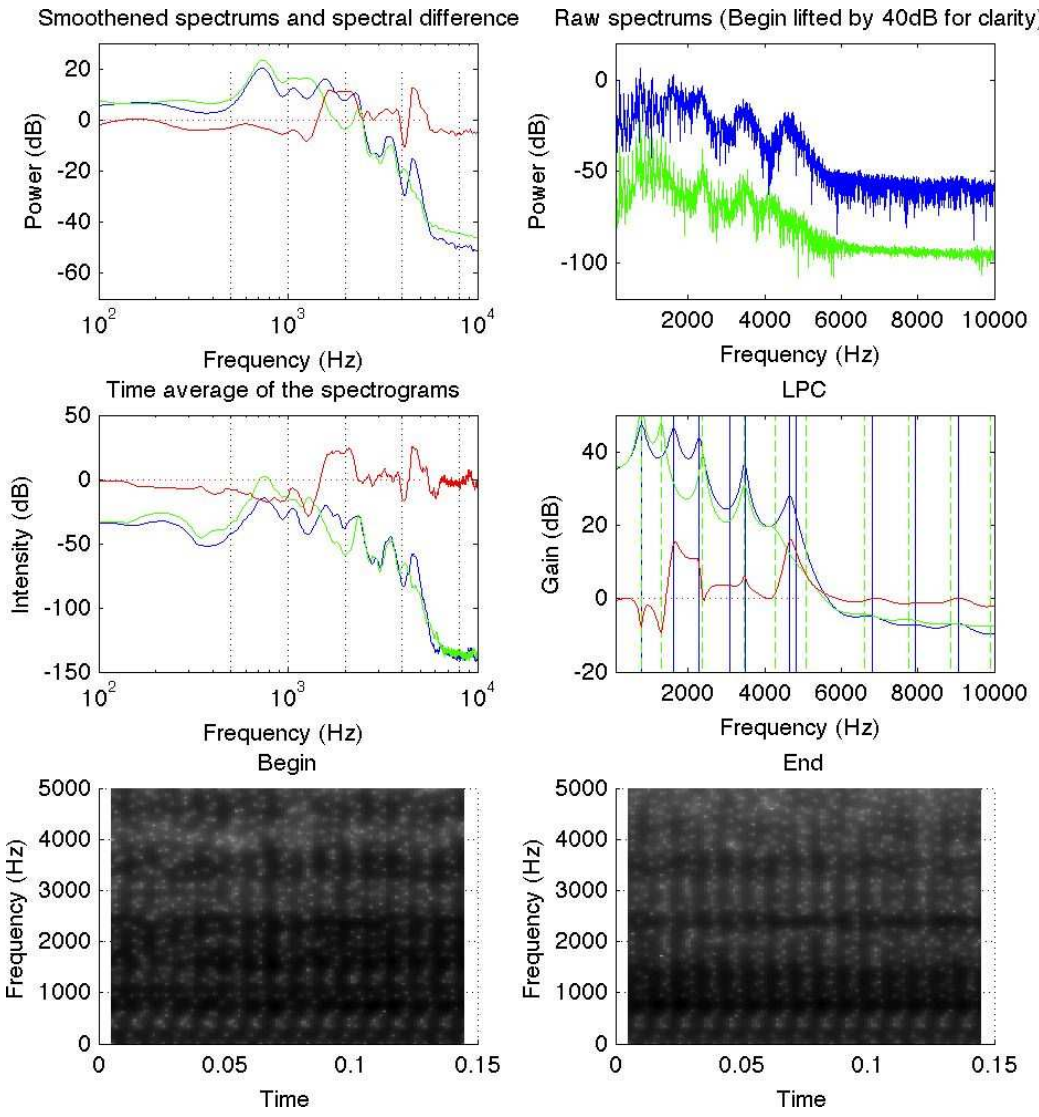


Figure A.7: [æ] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 4.1
Fundamental frequency in the begin sample: 106.3
Fundamental frequency in the end sample: 107.6
Formant distance: 517

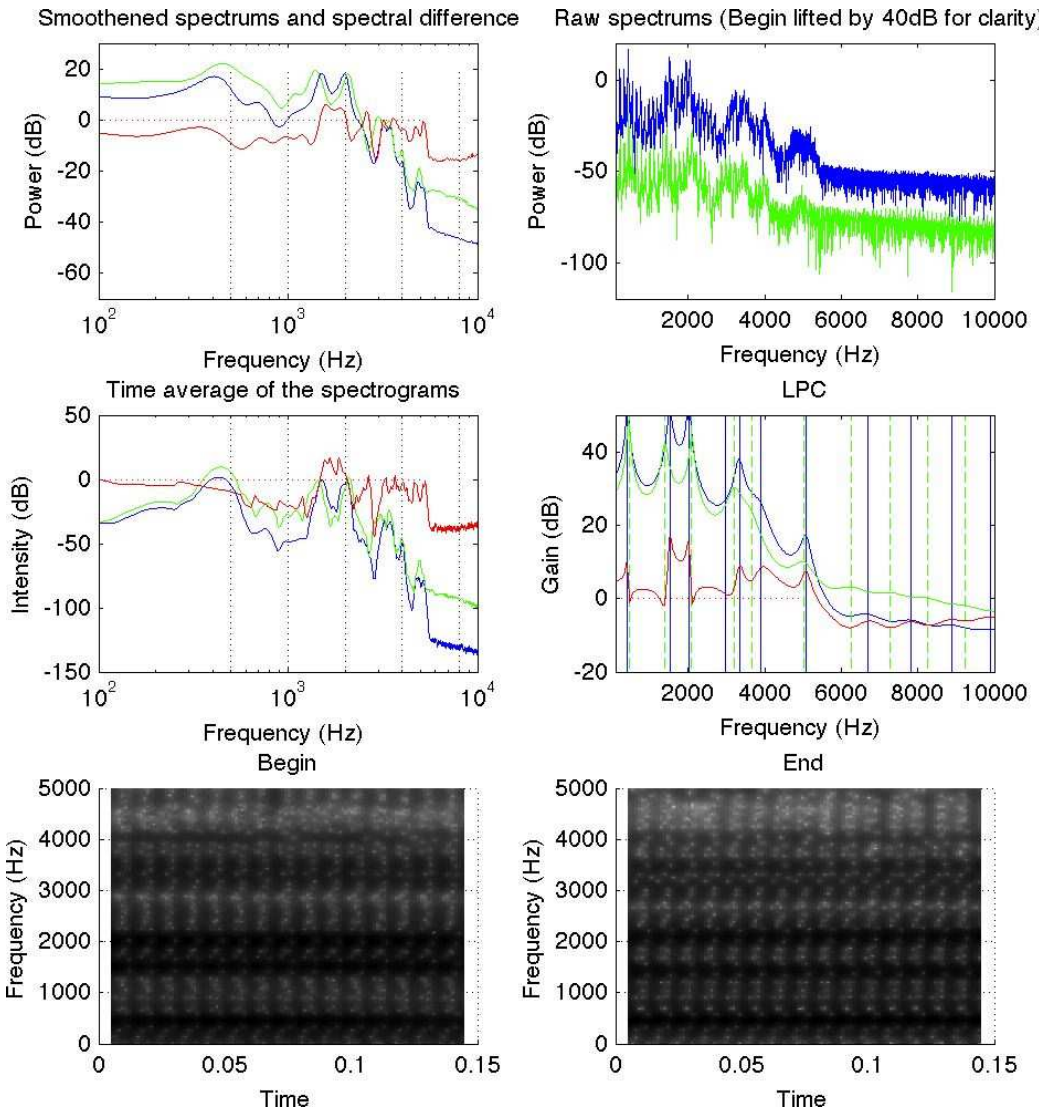


Figure A.8: [ø] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 2.2
Fundamental frequency in the begin sample: 107.6
Fundamental frequency in the end sample: 110.2
Formant distance: 285

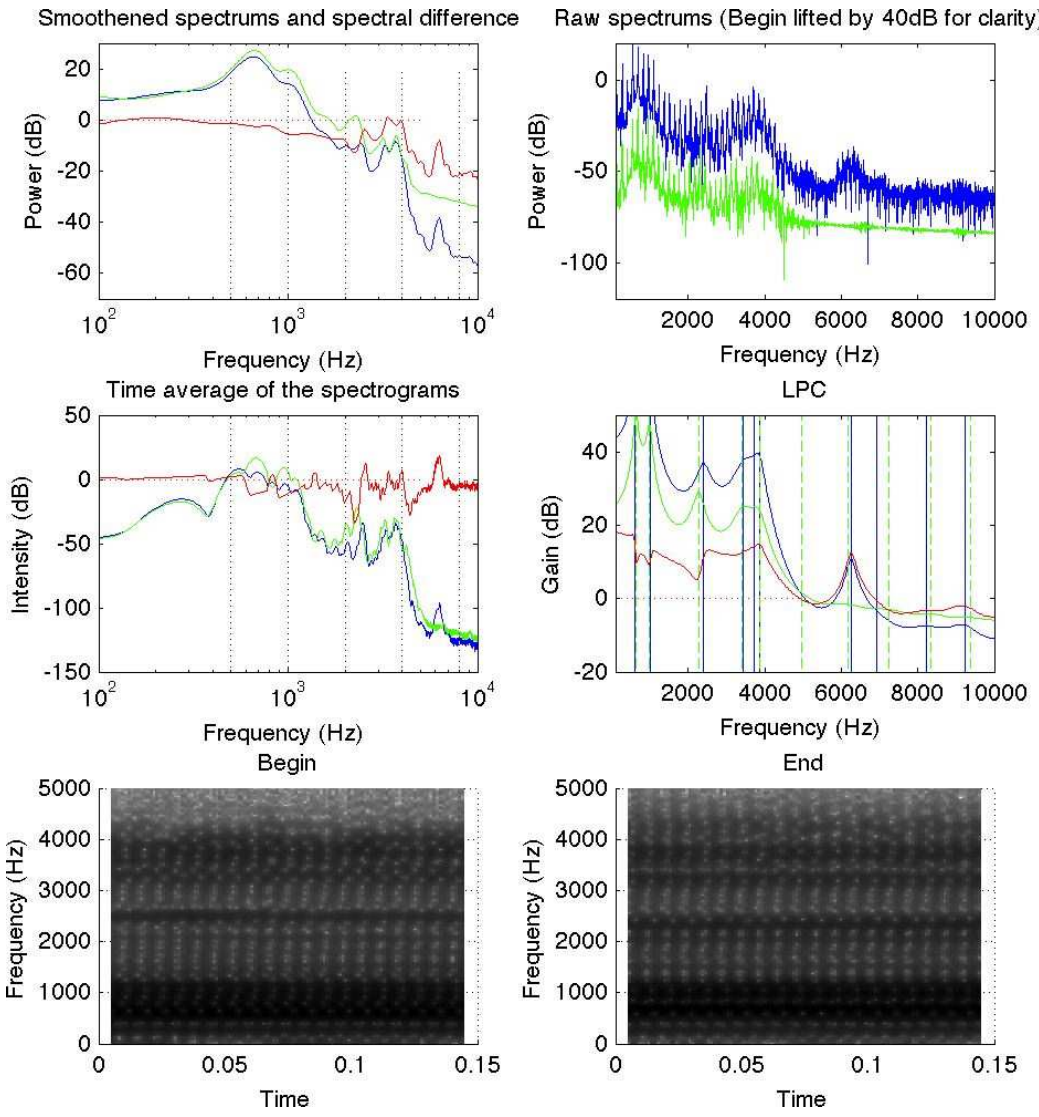


Figure A.9: [ɑ] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 3.2
Fundamental frequency in the begin sample: 137.0
Fundamental frequency in the end sample: 135.3
Formant distance: 134

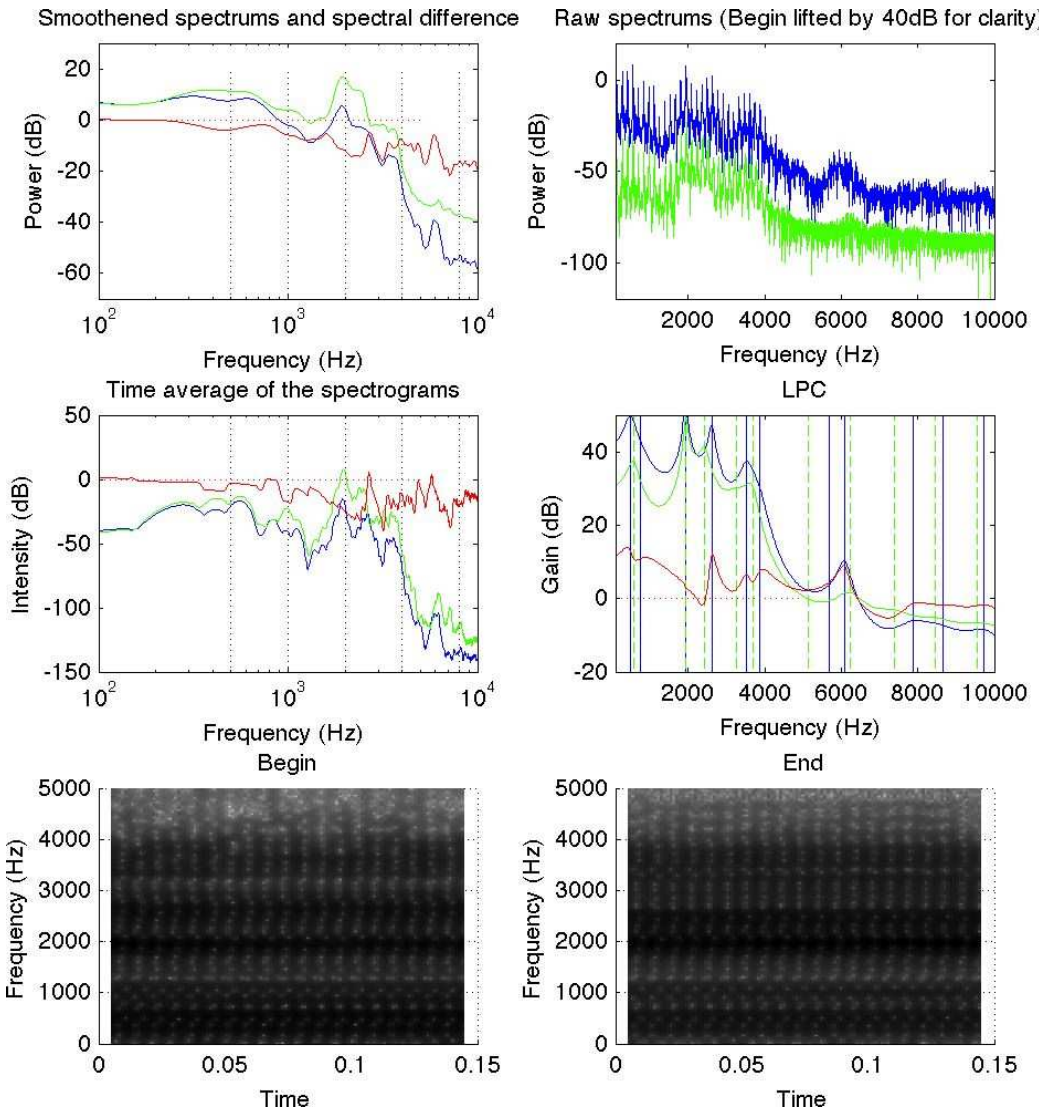


Figure A.10: [e] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 8.6
Fundamental frequency in the begin sample: 138.2
Fundamental frequency in the end sample: 139.6
Formant distance: 268

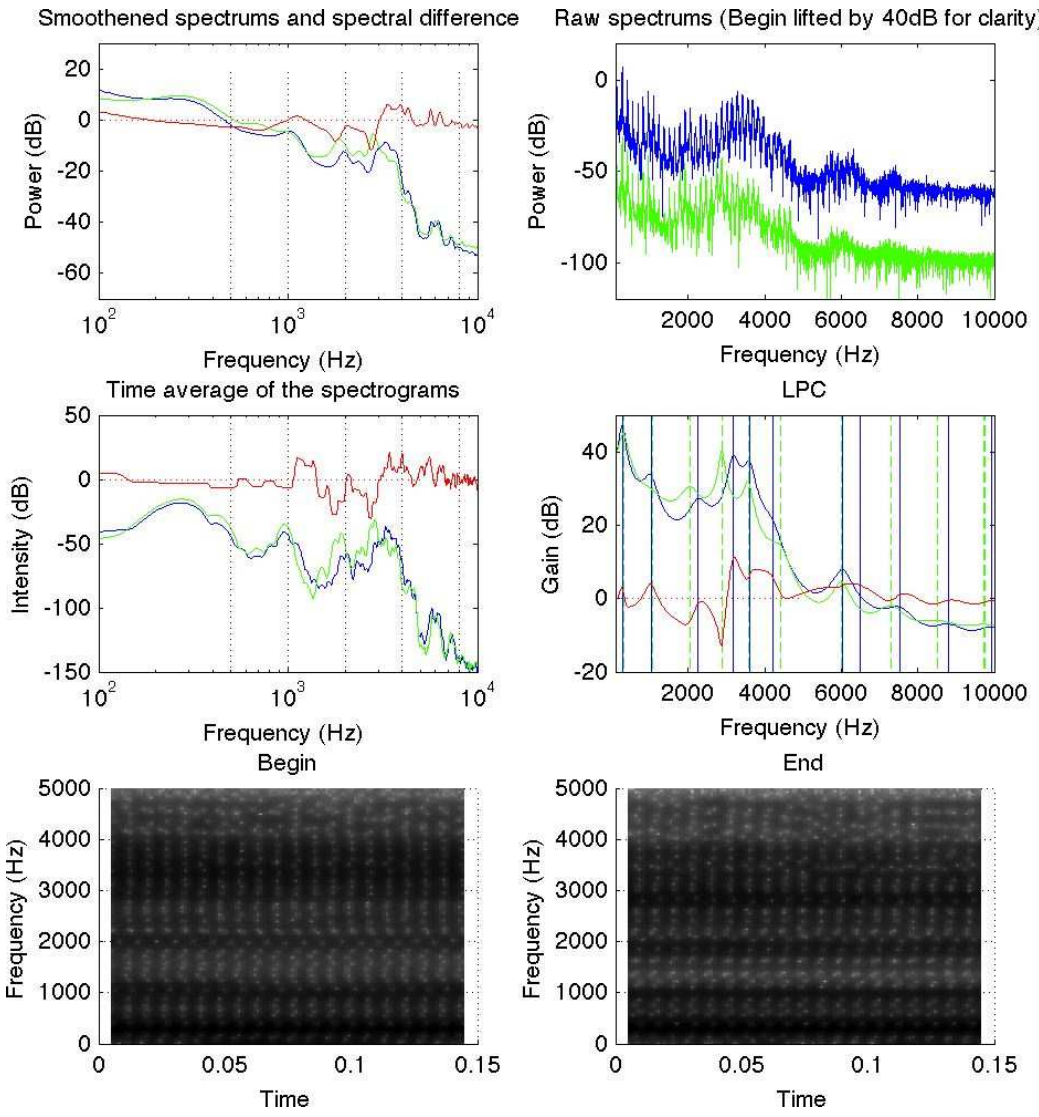


Figure A.11: [i] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 1.1
Fundamental frequency in the begin sample: 136.1
Fundamental frequency in the end sample: 136.5
Formant distance: 368

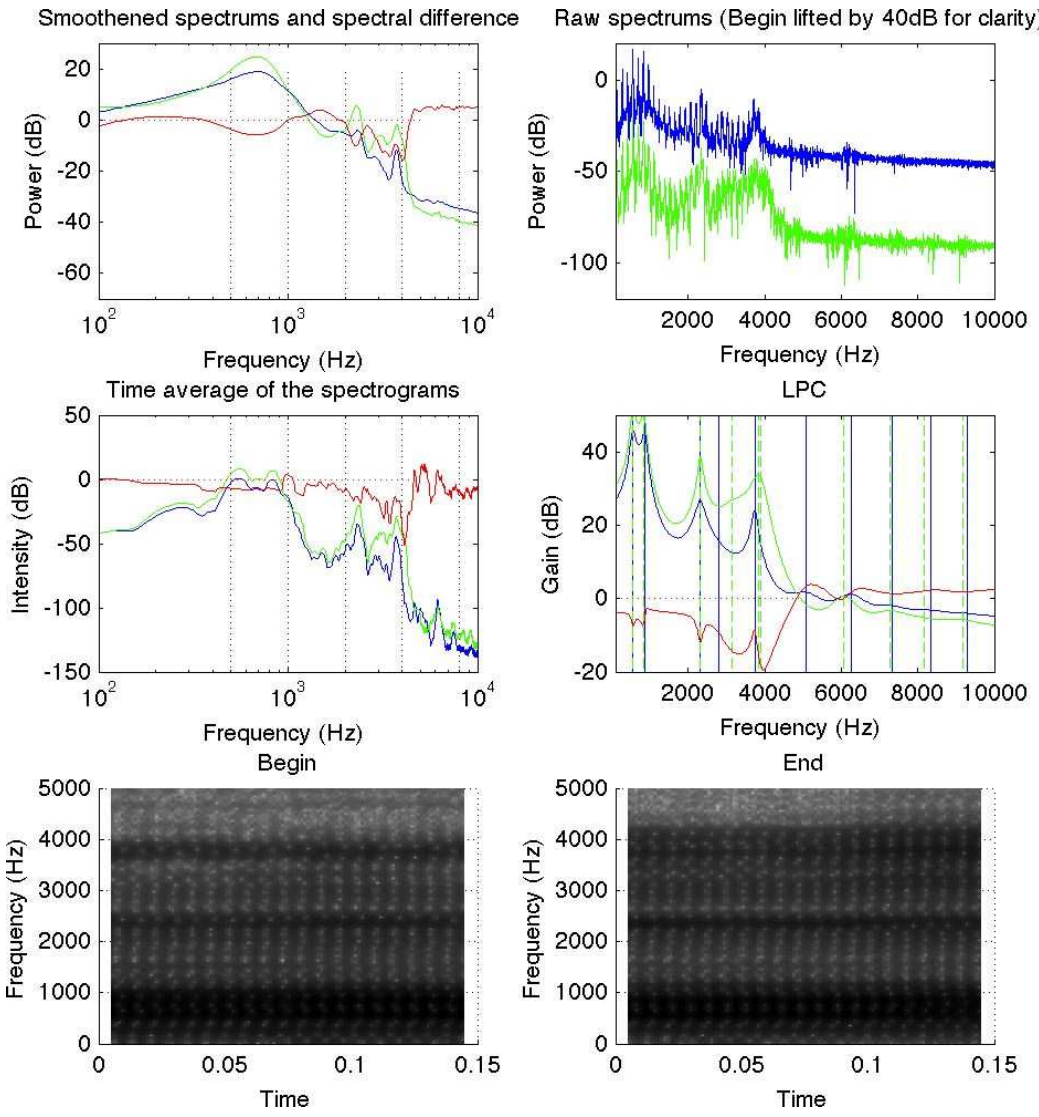


Figure A.12: [o] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 3.6
Fundamental frequency in the begin sample: 137.0
Fundamental frequency in the end sample: 138.2
Formant distance: 344

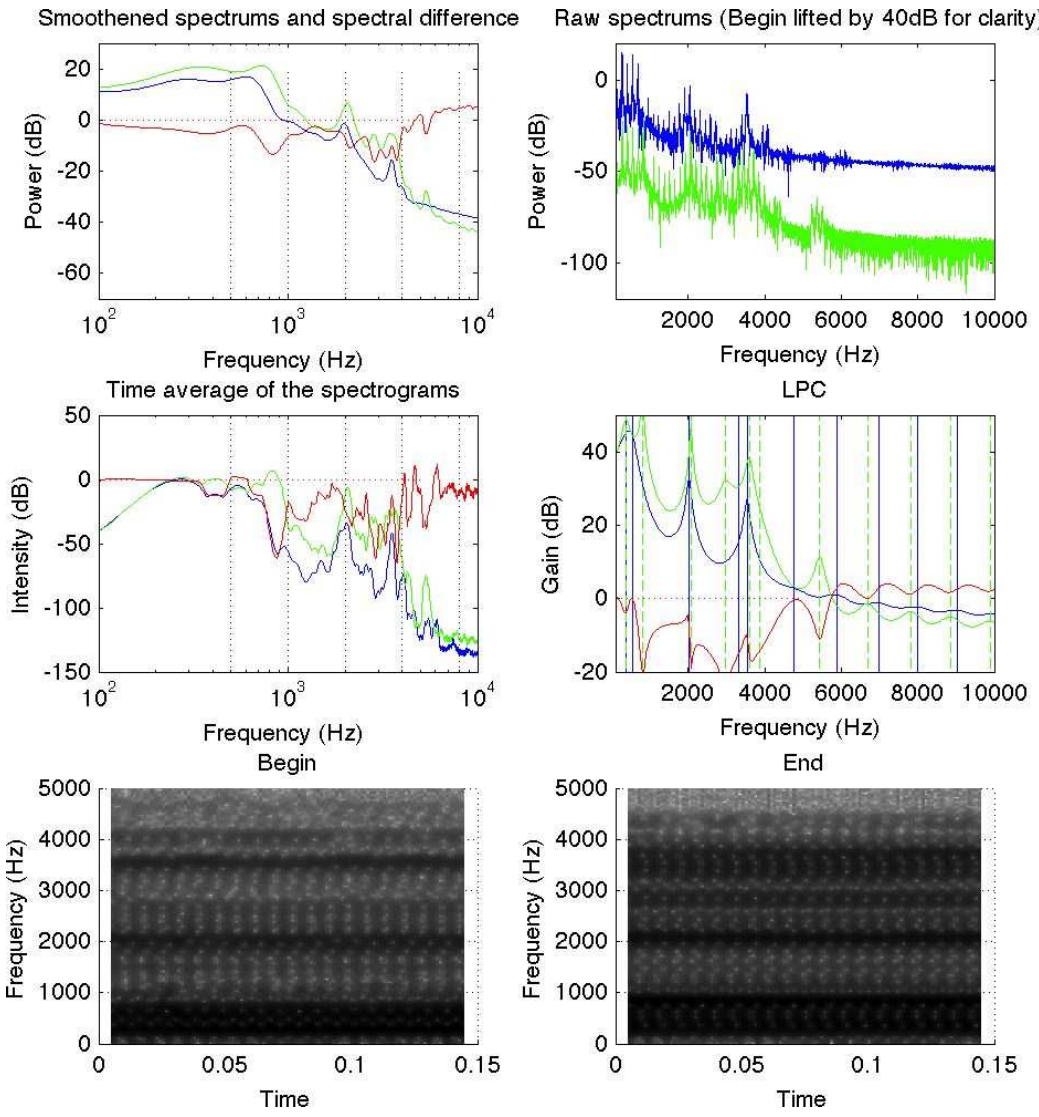


Figure A.13: [u] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 5.0
Fundamental frequency in the begin sample: 135.3
Fundamental frequency in the end sample: 137.0
Formant distance: 337

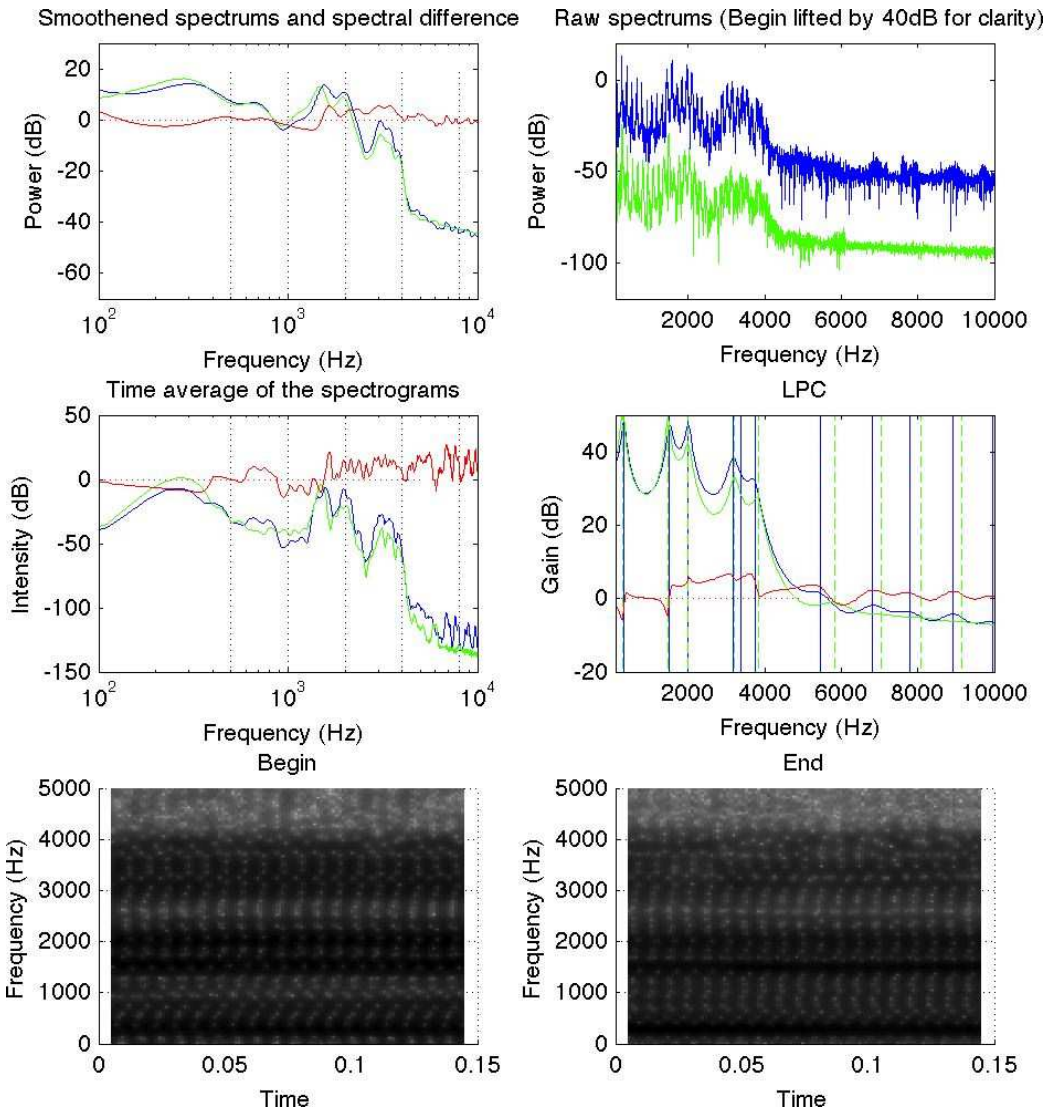


Figure A.14: [y] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: -0.1
Fundamental frequency in the begin sample: 130.9
Fundamental frequency in the end sample: 135.3
Formant distance: 64

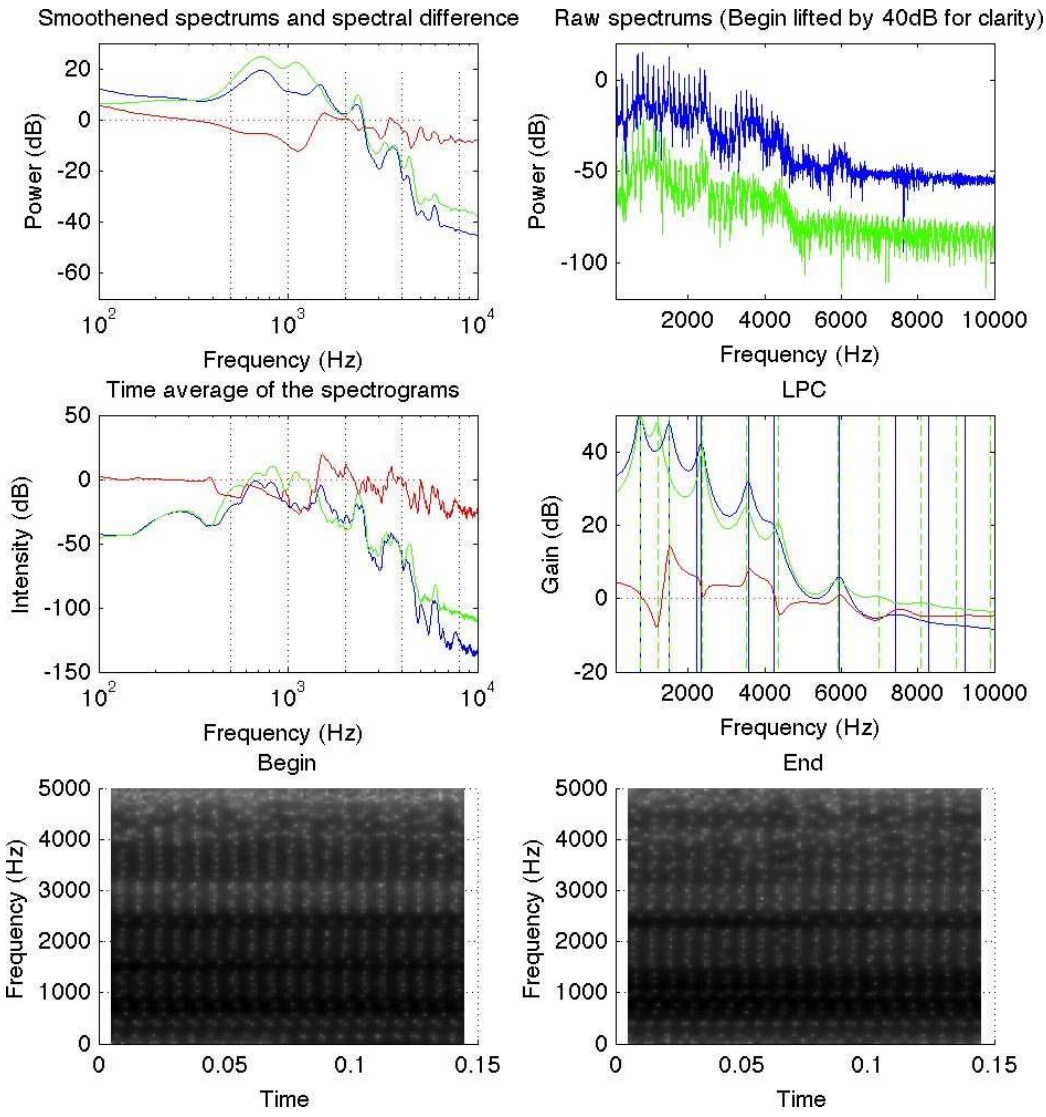


Figure A.15: [æ] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 4.8
Fundamental frequency in the begin sample: 134.9
Fundamental frequency in the end sample: 137.4
Formant distance: 151

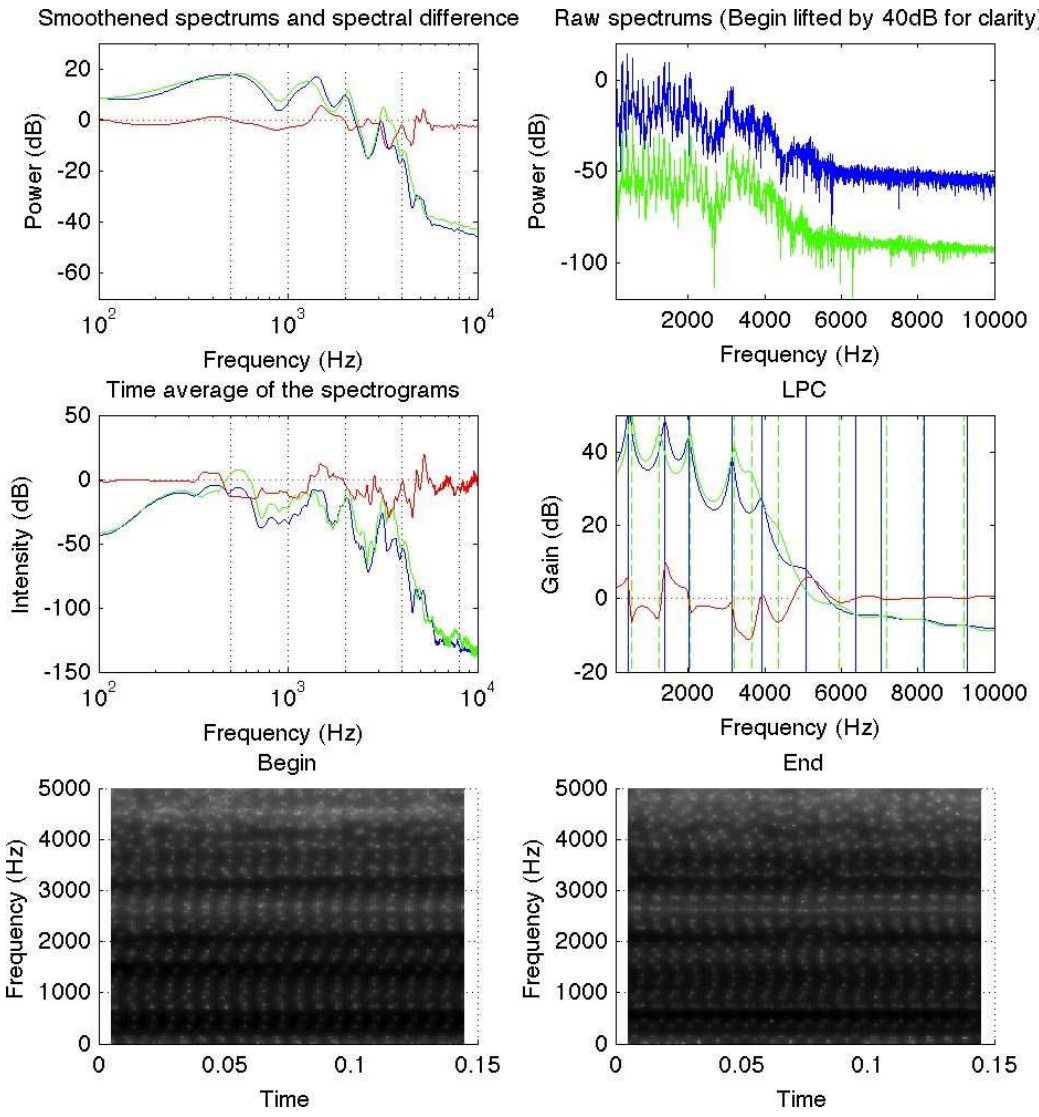


Figure A.16: [ø] target f_0 =137.5 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: 2.4
Fundamental frequency in the begin sample: 136.1
Fundamental frequency in the end sample: 137.0
Formant distance: 172

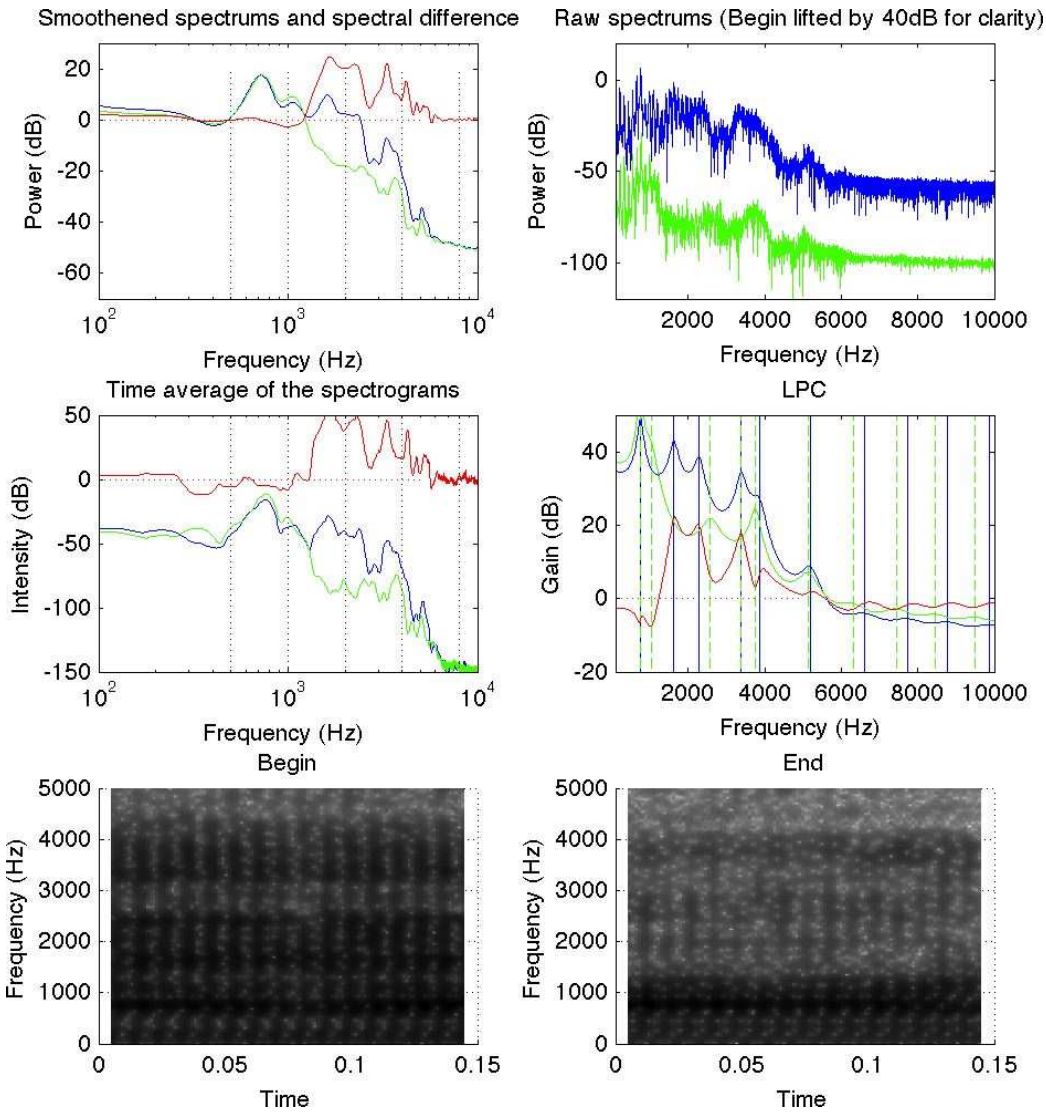


Figure A.17: [æ-ɑ] target f_0 =110 Hz, sequence: VIBE 1.8, duration 7.6 s.

Sound pressure level difference: -0.2
Fundamental frequency in the begin sample: 108.4
Fundamental frequency in the end sample: 109.4
Formant distance: 645

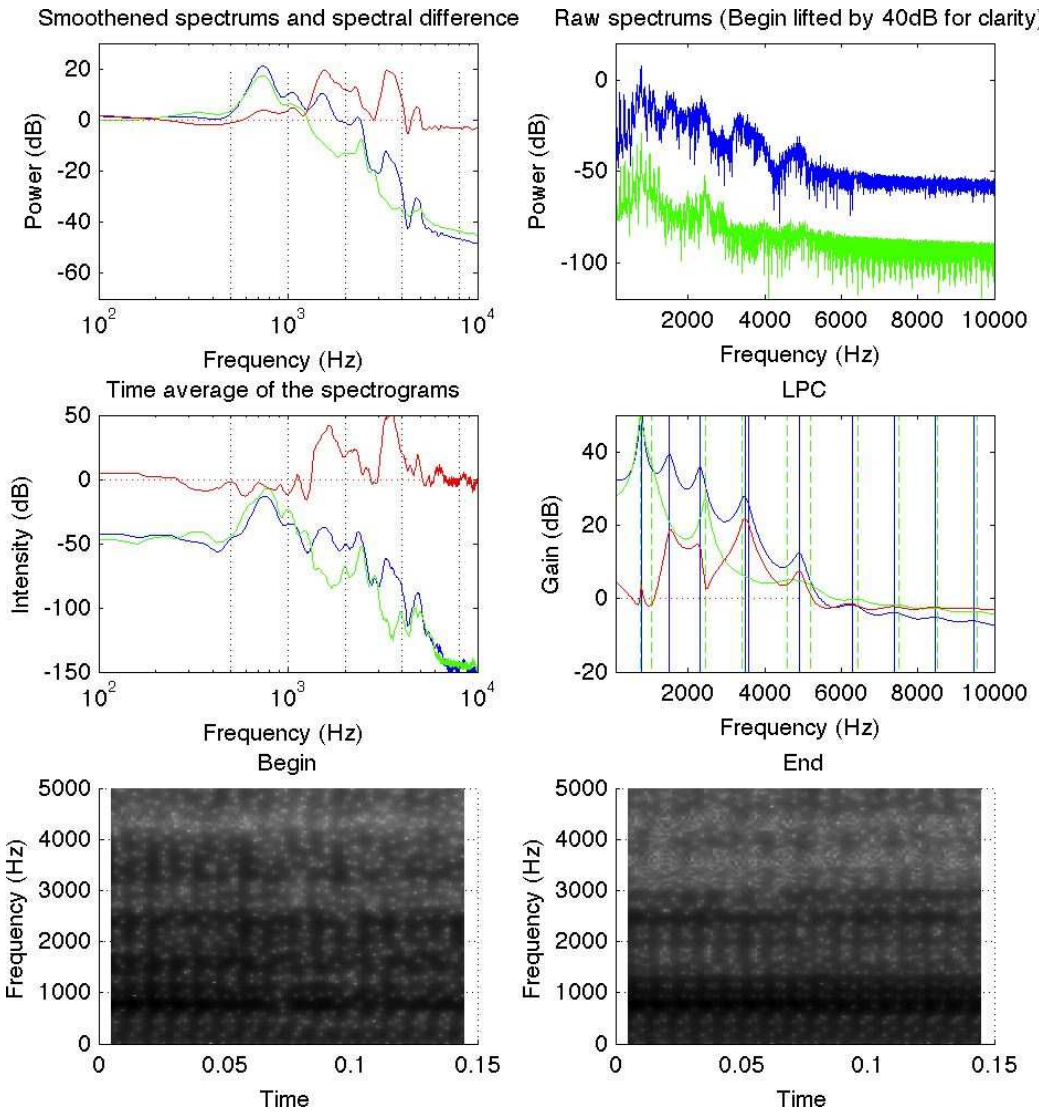


Figure A.18: [æ-ɑ] target f_0 =110 Hz, sequence: dynamic, duration 8 s.

Sound pressure level difference: 0.5
Fundamental frequency in the begin sample: 109.2
Fundamental frequency in the end sample: 110.5
Formant distance: 498

Appendix B

Deviation data on long vowel productions

Formant and f_0 extraction used the same algorithms and principles as that of the sound data from MRI (see Section 6.1 above). After extraction the variation in the samples was evaluated by computing the standard deviations (SD) of each of the acoustic measures with a moving 7.6 s window. Resulting SD data is illustrated on the following pages. In each Figure there are pannels where deviation data for f_0 and the formants F1-F4 are plotted as blue asterisks. Each of these five pannels has a red curve which plots the average deviation of the data set. In the sixth pannel is a plot of the product of f_0 and formants F1-F4 on a logarithmic scale.

The last page of this Appendix has histograms of the optimal sampling time distributions of the different measures. See Section 6.3 for a more detailed explanation.

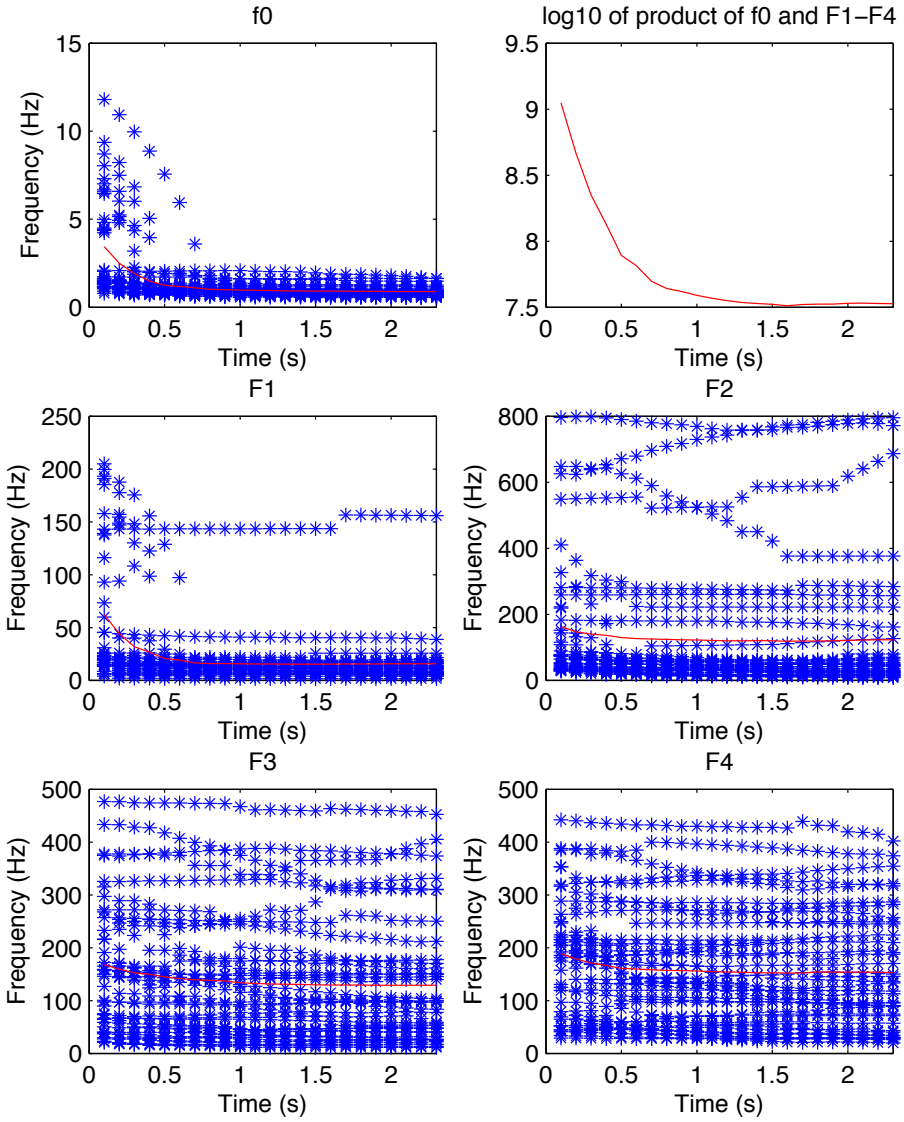


Figure B.1: Short and long phonations with both target f_0s.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4.

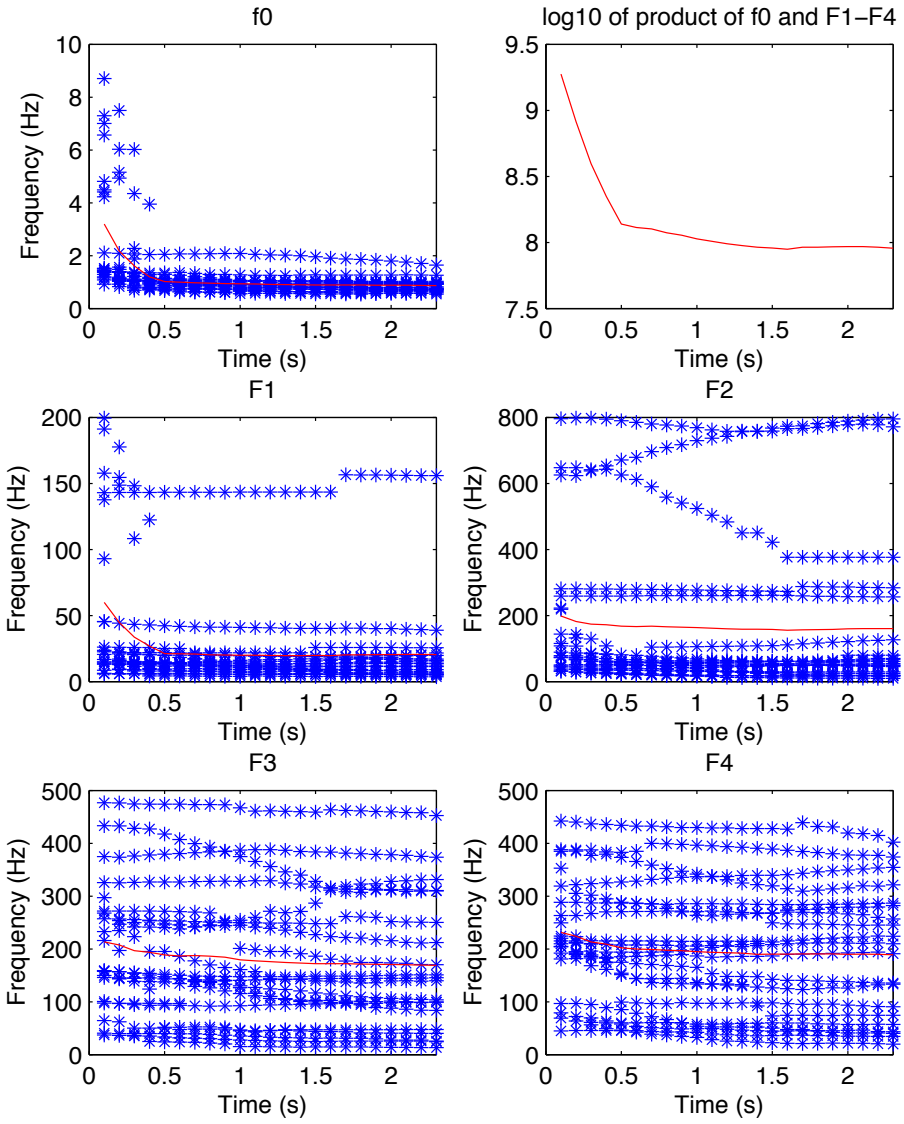


Figure B.2: Duration under 16 s.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4.

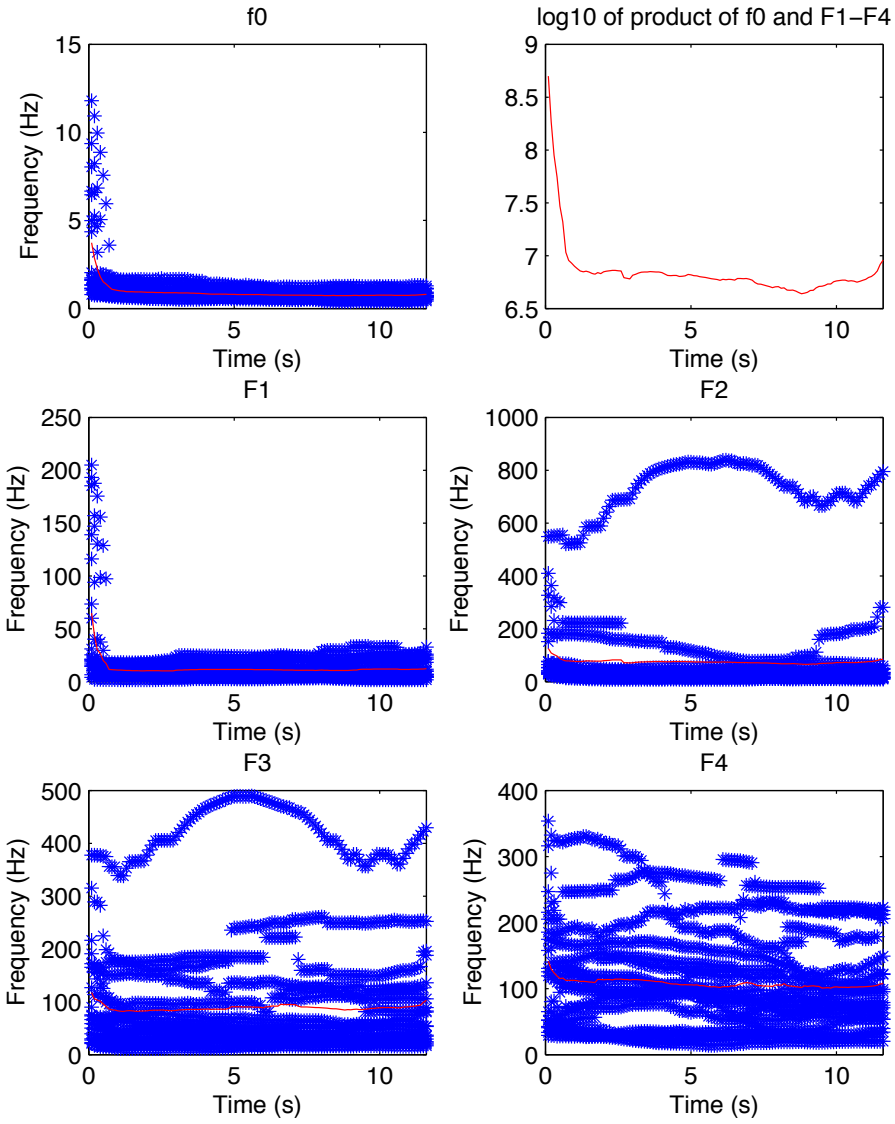


Figure B.3: Duration over 16 s.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4.

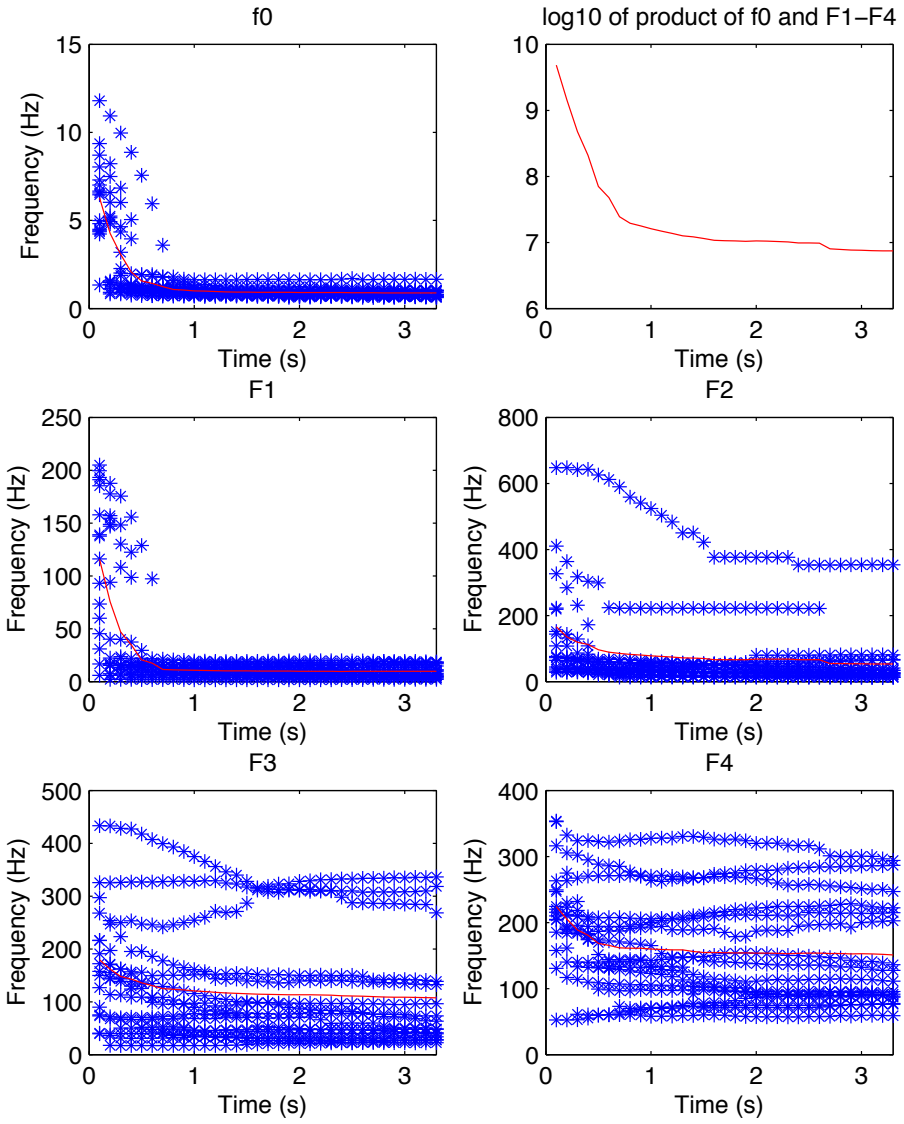


Figure B.4: Target f_0 = 110 Hz.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4.

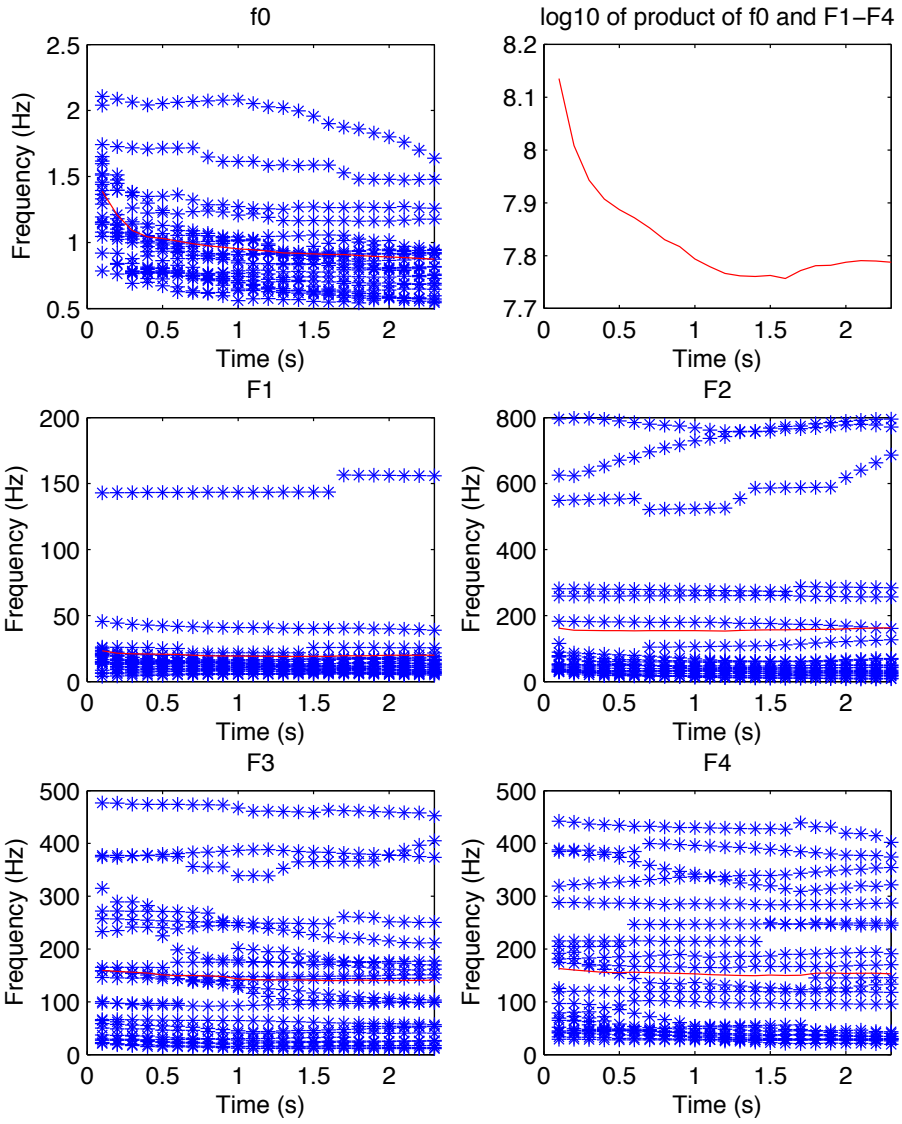


Figure B.5: Target f_0 = 137.5 Hz.: Standard deviations and base-10 logarithm of the product of standard deviations of f_0 and F1-F4.

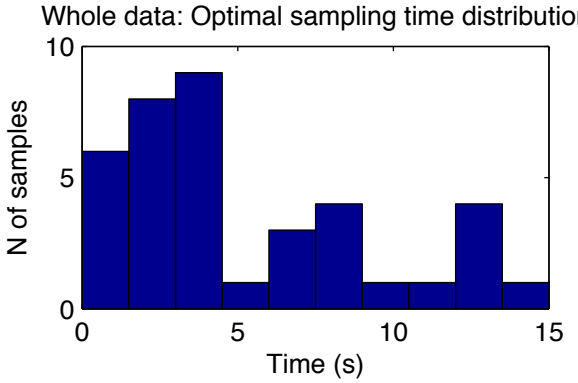


Figure B.6: Optimal sampling time distribution for the whole data.

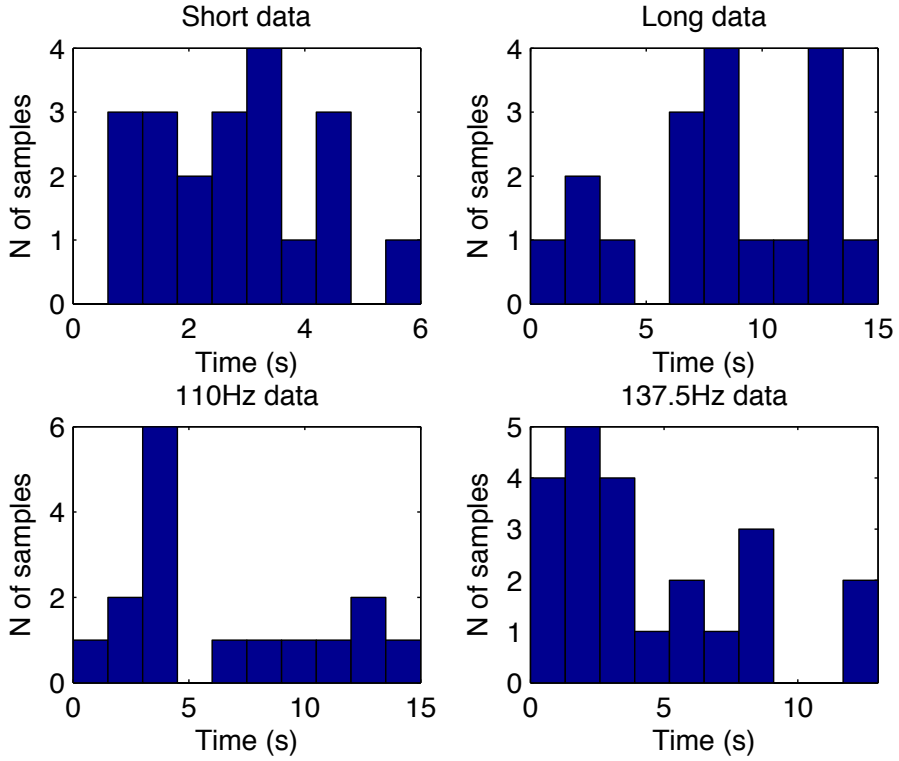


Figure B.7: Optimal sampling time distributions for different fractions of the data.