VILNIUS UNIVERSITY

Gražina Pyž

ANALYSIS AND SYNTHESIS OF LITHUANIAN PHONEME DYNAMIC SOUND MODELS

Doctoral Dissertation

Technological Sciences, Informatics Engineering (07T)

Vilnius, 2013

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2009–2013.

Scientific Supervisor

Assoc. Prof. Dr. Vytautas SLIVINSKAS (Lithuanian University of Educational Sciences, Technological Sciences, Informatics Engineering – 07 T).

Academic Consultant

Assoc. Prof. Dr. Virginija ŠIMONYTĖ (Lithuanian University of Educational Sciences, Technological Sciences, Informatics Engineering – 07 T).

VILNIAUS UNIVERSITETAS

Gražina Pyž

LIETUVIŠKŲ FONEMŲ DINAMINIŲ MODELIŲ ANALIZĖ IR SINTEZĖ

Daktaro disertacija

Technologijos mokslai, informatikos inžinerija (07T)

Vilnius, 2013

Disertacija rengta 2009–2013 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas

doc. dr. Vytautas SLIVINSKAS (Lietuvos edukologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Mokslinė konsultantė

doc. dr. Virginija ŠIMONYTĖ (Lietuvos edukologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Acknowledgments

I would like to express my thanks to all the people who have been in one way or another involved in the preparation of this thesis.

First of all, I would like to express my sincere gratitude to my scientific supervisor Prof. Vitautas Slivinskas and academic consultant Dr. Virginija Šimonytė for the continuous support of my research, for their patience, motivation, enthusiasm, and immense knowledge. I could not have imagined having better mentors for my doctoral study.

I want to express my big thanks to Vilnius University Institute of Mathematics and Informatics director Prof. Gintautas Dzemyda for granted conditions for doctoral studies. Special thanks to my thesis reviewers Prof. Kazys Kazlauskas and Dr. Gintautas Tamulevičius who read and provided constructive feedback regarding this dissertation. I would also like to thank the Lithuanian State Studies Foundation and Lithuanian Academy of Sciences for the financial support for doctoral studies.

Finally, I wish to thank all my relatives, friends and most of all for my husband Viktor for all their support and patience during this challenging period of my life.

Gražina Pyž

Abstract

Speech is the most natural way of human communication. Text-to-speech (TTS) problem arises in various applications: services for the hearing impaired, reading email aloud, reading text from e-book aloud, services for the people with speech disorders. A TTS system – a system that takes a sequence of words as input and converts it into speech.

In order to solve the problem of Lithuanian speech synthesis, it is necessary to develop mathematical models for Lithuanian speech sounds. In the dissertation proposed vowel and semivowel phoneme modelling framework is a part of this problem.

An assumption is made that a phoneme consists of the sum of components which could be generated by properly chosen formant synthesizer parameters. The second order quasipolynomial is chosen as the component model in time domain. Multiple input and single output systems (MISO), whose inputs are sequences of amplitude modulated impulses, are used for sound modelling. The frequencies, damping factors, amplitudes and phases are parameters of these systems. The synthesizing consists of the following two stages: 1) phonemic synthesizer parameter estimation from the characteristic period data, 2) determination of the exciting input impulse periods and amplitudes.

In order to estimate the synthesizer parameters, the real sound signals that are expanded into components by the inverse fast Fourier transform method are used. The new parameter estimation algorithm for convoluted data, based on Levenberg-Marquardt approach, has been derived and its stepwise form presented.

In order to obtain more natural sounding of the synthesized speech, it is important to use not only high-order models, but complex input sequence scenarios as well. The input sequence of each phoneme has been described by three parabolas. The first parabola characterized the slowdown growth of the

component, the second parabola described the main time region, and the third one characterized the slowdown decreasing of the component. Transition from one vowel to another vowel was achieved by changing excitation impulse amplitudes by the arctangent law.

The synthesized sounds have been audio tested, and the Fourier transforms of the real data and output of the MISO model have been compared. It was impossible to distinguish between the real and simulated data. The magnitude and phase responses have only shown small differences. The practical results have shown that the synthesized sounds of this method are sufficiently natural, pleasantly sounding.

List of Figures

List of Tables

Notations

Symbols

$s(n)$	*The phoneme signal*
f_0	*The fundamental frequency*
f_s	*The sampling frequency*
$\boldsymbol{R}$	*The covariance matrix*
γ_k	*The eigenvalue of the covariance matrix*
$\widehat{\boldsymbol{P}}_{MU}$	*The MUSIC spectral function*
$\boldsymbol{h}_k$	*The impulse response of the k-th SISO system*
$\boldsymbol{u}_k$	*The input of the k-th SISO system*
$\boldsymbol{y_k}$	*The output of the k-th SISO system*
$q(t)$	*The quasipolynomial*
f	*The frequency*
Ω	*The angular frequency*
φ_i	*The phase*
a_i	*The amplitude of continuous -time signal*
A_i	*The amplitude of a discrete-time signal*
λ	*The damping factor of continuous -time signal*
Λ	*The damping factor of a discrete-time signal*

$\boldsymbol{\Phi}$	*The convoluted basis signal matrix*
$\boldsymbol{\Psi}$	*The standard basis signal matrix*
$\boldsymbol{\theta}$	*The parameter vector*
T	*The period*

Abbreviations and Acronyms

DFT	*Discrete Fourier transform*
IPA	*International Phonetic Alphabet*
LPC	*Linear Predictive Coding*
MISO	*Multiple-Input and Single-Output*
MUSIC	*MUltiple SIgnal Classification*
RMSE	*Root-Mean-Square Error*
SISO	*Single-Input and Single-Output*
TTS	*Text-to-Speech*

Contents

1

Introduction

1.1. Research Context and Challenges

A Text-to-speech (TTS) system is defined as a system that takes a sequence of words as input and converts it into speech (SIL, 2004). The speech synthesizer can be useful in many cases. TSS system can read aloud any texts from web pages, navigation, translation and other applications. It helps with pronunciation and learning foreign languages, promoting listening skills. The speech synthesizer allows multi-tasking so that attention can be given to reading materials when time would otherwise not permit. It helps people with reading challenges, or visual impairment.

Construction of speech synthesizer is a very complex task. Researchers are trying to automate speech synthesis. Yet there is no automatic Lithuanian TTS system equivalent to human speech. The commercial TTS systems have not yet supported Lithuanian language. The problem of developing Lithuanian synthesizer arises. There exists a Lithuanian synthesizer developed by P. Kasparaitis (P. Kasparaitis, 2001). It is based on concatenation speech

synthesis type. Concatenation synthesis relies on speech sounds recorded in advanced database. One of the main drawbacks of concatenation synthesis is that the database has to be sufficiently large. That, however, requires extensive computer resources. If a word is not in the database, then it could not be synthesized. The synthesized speech quality does not achieve the natural speech quality since glitches occur on the concatenation boundaries. Formant synthesis does not require a sound database. Formant synthesizers have advantages against the concatenative ones. The speech produced by them can be sufficiently intelligible even at high speed. They can control prosody aspects of the synthesized speech (intonation, rhythm, stress). The main drawback of formant synthesis is that the sounds obtained by this synthesis type sound unnaturally, robot-like. In this work an assumption is made that the models of formant synthesizer are too simple. In order to reduce synthetic sounding, it is necessary to develop new mathematical models for speech sounds. The vowel and semivowel phoneme models are a part of this problem.

1.2. Problem Statement

The sounds obtained by formant synthesis type sound unnaturally, robot-like. In order to reduce synthetic sounding, it is necessary to develop new mathematical models for speech sounds, which could be used as a base of speech synthesizer.

1.3. Object of Research

The research object of the dissertation is Lithuanian vowel and semivowel phoneme models.

1.4. The Objective and Tasks of the Research

The objective of the thesis is to develop Lithuanian speech vowel and semivowel phoneme dynamic models, and create transition between phonemes in order to join these models.

In order to achieve the objective, the following tasks are stated:

- to acquaint with speech production apparatus, main speech synthesis methods and the existing text-to-speech systems of Lithuanian and other languages.
- to analyse main characteristics of Lithuanian speech vowel and semivowel sounds.
- to ascertain what models are suited best for Lithuanian speech sound description.
- to develop mathematical models of Lithuanian speech vowel and semivowel phonemes.
- to create transitions between phonemes in order to join vowel and semivowel models.
- to evaluate the proposed models accuracy experimentally.

1.5. Methodology of Research

- Digital signal processing,
- System theory,
- Optimization methods,
- Matrix algebra,
- Mathematical statistics,
- Programming in Matlab environment,
- Programming in C# language.

1.6. Scientific Novelty

- For vowel and semivowel phonemes modelling MISO system whose impulse response of each channel is described as a third order quasipolynomial and input amplitude impulse vary in time is proposed.
- A new parameter estimation algorithm for convoluted data, based on Levenberg-Marquardt approach, has been derived.
- A new fundamental frequency refining algorithm is proposed.
- A new method that allows one to select the representative period automatically is given.
- The transitions between vowel and semivowel phoneme models have been derived.

The advantage of my developed phoneme modelling framework is that anyone can use it and it can synthesize any phoneme of vowel and semivowel for any speaker.

1.7. Practical Significance of the Results

The proposed vowel and semivowel phoneme models can be used for developing a TTS formant synthesizer. The phoneme models can also be adapted to other similar problems, for example, treating language disorders, helping with pronunciation and learning of foreign languages.

1.8. Defended Propositions

1) For vowel and semivowel phoneme modelling a discrete time linear stationary system with multiple-input and single-output (MISO) is used.

Impulse response of each MISO system channel is described as a third order quasipolynomial.

2) In order to obtain more natural sounding of the synthesized speech, it is important to use not only high-order models, but complex input sequence scenarios as well.

3) The vowel and semivowel synthesis quality is sufficiently good.

4) The word consisting of vowels and semivowels obtained with the proposed synthesis methods is enough and it is difficult to distinguish it from the real one. The quality of the synthesized sound was significantly improved due to input transitions.

1.9. Approbation and Publications of the Research

The main results of the dissertation were published in 6 articles in the periodical scientific publications. The main results of the work have been presented and discussed at 21 national and international conferences.

International conferences

1. The 5th International Conference Mechatronic Systems and Materials, October 22 - 25, 2009, Vilnius, Lithuania.
2. International Conference of Young Scientists, April 29-30, 2010, Šiauliai, Lithuania.
3. The 6th International Conference on Electrical and Control Technologies ECT-2011, May 5-6, 2011, Kaunas, Lithuania.
4. The 2nd International Doctoral Consortium Informatics and Informatics Engineering Education Research: Methodologies, Methods, and Practice, November 30-December 4, 2011, Druskininkai, Lithuania.

5. The 7th International Conference on Electrical and Control Technologies ECT-2012, May 3-4, 2012, Kaunas, Lithuania.
6. The 8th Joint European Summer School on Technology Enhanced Learning, May 21-25, 2012, Estoril, Portugal.
7. The 13th International conference Teaching Mathematics: Retrospective and Perspectives, 30 May – 1 June, 2012, Tartu, Estonia.
8. The 16th International Conference ELECTRONICS'2012 18-20 June, 2012, Palanga, Lithuania.
9. The 2nd International Conference Music and Technologies, 8-10 November, 2012, Kaunas, Lithuania.
10. The 3rd International Doctoral Consortium Informatics and Informatics Engineering Education Research: Methodologies, Methods, and Practice, December 3-7, 2012, Druskininkai, Lithuania.
11. The 8th International Conference on Electrical and Control Technologies ECT-2013, May 2-3, 2013, Kaunas, Lithuania.
12. The 14th International conference Teaching Mathematics: Retrospective and Perspectives, May 9–11, 2013, Jelgava, Latvia.

Regional conferences

1. Lietuvos matematikų draugijos 50-oji konferencija, Vilnius: MII, 2009 m. birželio 18–19 d.
2. 1-oji jaunųjų mokslininkų konferencija „Fizinių ir technologijos mokslų tarpdalykiniai tyrimai“, Vilnius: LMA, 2011 m. vasario 8 d.
3. Lietuvos matematikų draugijos 52-oji konferencija, Vilnius: LKA, 2011 m. birželio 16-17 d.

4. 15-oji mokslinė kompiuterininkų konferencija „Kompiuterininkų dienos 2011“, Klaipėda: KU, 2011 m. rugsėjo 22–24 d.
5. 2-oji jaunųjų mokslininkų konferencija Fizinių ir technologijos mokslų tarpdalykiniai tyrimai, Vilnius: LMA, 2012 m. vasario 14 d.
6. Lietuvos matematikų draugijos 53-oji konferencija, Klaipėda: KU, 2012 m. birželio 11-12 d.
7. 3-oji jaunųjų mokslininkų konferencija Fizinių ir technologijos mokslų tarpdalykiniai tyrimai, Vilnius: LMA, 2013 m. vasario 12 d.
8. Lietuvos matematikų draugijos 54-oji konferencija, Vilnius: LEU, 2013 m. birželio 19-20 d.
9. 16-oji mokslinė kompiuterininkų konferencija „Kompiuterininkų dienos 2011“, Šiauliai: ŠU, 2013 m. rugsėjo 19–21 d.

1.10. Outline of the Dissertation

The dissertation consists of 5 chapters, references and appendices. The total scope of the dissertation without appendices – 114 pages containing 78 formulas, 47 pictures and 19 tables.

The Introduction (Chapter 1) reveals research context and challenges, describes the problem statement, the object of research, the tasks and objective of the dissertation, methodology of research, presents scientific novelty, practical significance of results, defends propositions and approbation of obtained results.

In Chapter 2 an overview of Lithuanian speech engineering is given. Detailed information about Lithuanian speech phonemes and diphthongs is presented.

In Chapter 3 Lithuanian vowel and semivowel phoneme modelling framework is submitted. Within this framework two synthesis methods are

proposed: harmonic and formant.

Chapter 4 provides the results of experimental researches.

Conclusions (Chapter 5) present the main conclusions of the dissertation.

Appendices present a list of Lithuanian phonemes with the examples of their usage, the plots of the vowel and semivowel phoneme signals.

2

Fundamentals of speech synthesis

This chapter provides an overview of Lithuanian speech engineering. Speech synthesis is one of its parts. Two speech synthesis types are presented. Their advantages and disadvantages are listed. The MUSIC method for the estimation of the signal fundamental frequency is submitted. An overview of Lithuanian speech phonemes and diphthongs is given. The differences between the fundamental frequency values of unstressed and stressed vowel and semivowel phonemes have been revealed.

2.1. Speech engineering in Lithuania

Speech engineering is a popular field of engineering in Lithuania. Much effort is given by Lithuanian scientists and engineers for developing digital technologies of Lithuanian speech processing. An overview of speech engineering in Lithuania at the end of the twentieth century is given in (Lipeikienė and Lipeika, 1998). The dominating field of Lithuanian speech

engineering is speech recognition. Speech recognition is a process that converts spoken words and phrases to a computer-readable format. Speech recognition is mostly developed for the languages of big countries especially for English language. There exists a number of programs (Dragon Naturally Speaking, Speech Recognition in Windows 7, Dragon Dictate for iPhone/iPad, etc.) that are intended for the use with English language. The languages of small countries cannot boast of such a great attention, nevertheless the local researchers give much effort to developing recognizers of the local languages.

Lithuanian speech recognition is one of the Lithuanian speech processing problems taking considerable attention of Lithuanian researchers. Research groups work in Vilnius at the Institute of Mathematics and Informatics and the Faculty of Mathematics and Informatics of Vilnius University, in Kaunas at Vytautas Magnus University, Kaunas University of Technology and Vilnius University Kaunas Faculty of Humanities. Some of the problems analysed by researchers of the Institute of Mathematics and Informatics are as follows: development of isolated word speech recognition system (Lipeika et al., 2002), application of dynamic programming for word endpoint detection in isolated word recognition (Lipeika and Lipeikienė, 2003), creating a framework for choosing a set of syllables and phonemes for Lithuanian speech recognition (Laurinčiukaitė and Lipeika, 2007), using the formant features in the dynamic time warping based recognition of isolated Words (Lipeika and Lipeikienė, 2008; Lipeika, 2010), quality estimation of speech recognition features (Lileikytė and Telksnys, 2011; Lileikytė and Telksnys, 2012), speaker recognition by voice (Kamarauskas, 2009), development of isolated word recognition systems (Tamulevičius, 2008). Scientists from Vilnius University the Faculty of Mathematics and Informatics and Forensic Science Centre of Lithuania investigate speaker recognition problems (Bastys et al., 2010), evaluation of effectiveness of different methods in speaker recognition (Šalna and Kamarauskas, 2010), Lithuanian speech recognition using the English

recognizer (Kasparaitis, 2008). Researchers from Vilnius Gediminas Technical University deal with control of robots by voice (Navakauskas and Paulikas, 2006), development of biometric systems for person recognition (Ivanovas and Navakauskas, 2010), development and implementation of means for word duration signal processing (Ivanovas, 2012; Tamulevičius et al., 2010). Scientists from Vytautas Magnus University deal with building medium-vocabulary isolated-word Lithuanian HMM speech recognition system (Raškinis and Raškinienė, 2003), modelling phone duration of Lithuanian by classification and regression trees (Norkevičius and Raškinis, 2008), investigating hidden Markov model modifications for large vocabulary continuous speech recognition (Šilingas and Telksnys, 2004), analysis of factors influencing accuracy of speech recognition (Čeidaitė and Telksnys, 2010). Researchers at Kaunas University of Technology and Vilnius University Kaunas Faculty of Humanities investigate foreign languages models for Lithuanian speech recognition (Maskeliūnas et al., 2009), the improvement of voice command recognition accuracy (Maskeliūnas, et al., 2011), deal with implementation of hierarchical phoneme classification approach on LTDIGITS corpora (Driaunys et al., 2009), consider control of computer and electric devices by voice (Rudžionis et al., 2008).

Lithuanian speech synthesis is a part of Lithuanian speech digital processing area that attracts considerable attention. Significant results in Lithuanian speech synthesis have been achieved by P. Kasparaitis in collaboration with experts of Lithuanian language (The MBROLA Project). The synthesizer developed by him is based on concatenation synthesis method. This method exploits Lithuanian speech corpora developed in advance. T. Anbinderis investigates one of the constituent parts of speech synthesis – automatic stressing of a text (Anbinderis, 2010). What concerns Lithuanian speech, formant synthesis has not yet attracted much attention of researchers. Problems related to developing Lithuanian speech formant synthesizers are

considered in (Ringys and Slivinskas, 2009; Ringys and Slivinskas, 2010).

Speech animation problems also attract attention of Lithuanian researchers. One of such problems is Lithuanian phoneme visualization. A methodology of such visualization is proposed in (Mažonavičiūtė and Baušys, 2011).

Speech analysis is another large class of speech processing problems. The paper (Balbonas and Daunys, 2007) can be mentioned as an example of such analysis. Some researchers try to use Wienerclasssystems for speech signal prediction (Ivanovas and Navakauskas, 2011).

Other fields (some of them are closely related with speech recognition) of Lithuanian language and speech engineering are noisy speech intelligibility enhancement (Kazlauskas, 1999), intelligent extraction of an internal signal in a Wiener System (Kazlauskas and Pupeikis, 2013), transcribing of the Lithuanian text (Kasparaitis, 1999; Skripkauskas and Telksnys, 2006), automatic stressing of the Lithuanian text (Kasparaitis, 2000; Anbinderis, 2010), coding and transmission of voice signals (Kajackas and Anskaitis, 2009), the Lithuanian language machine translation (Šveikauskienė, 2005).

2.2. Speech synthesis types

Researchers have been showing interest in speech synthesis for a long time (see, e. g., (Hopcroft and Ullman, 1979; Holmes and Holmes, 2001; Slivinskas and Šimonytė, 2007)). The best known commercial TTS systems are Bell Labs TTS and Festival developed at University of Edinburgh. The construction of a model for segmental duration in German is considered in the paper (Mobius and van Santen,1996). This model has been implemented in the German version of the Bell Labs text-to-speech system. The goal of the paper (Mobius and Von Santen,1996) was to analyse and model durational patterns of natural speech in order to achieve an improved naturalness of synthetic speech. Although many results have been achieved, this field still remains important.

There exist two main speech signal synthesis types: concatenative synthesis and formant synthesis (Donovan, 1996; Frolov A. and Frolov G., 2003). Synthesized speech sounds are created using concatenation of pieces of recorded speech stored in a database in concatenative synthesis. Formant synthesizers do not use any recorded sounds. A speech signal is modelled as an output of a linear filter and is described by a mathematical model with a finite number of parameters.

Both speech synthesis types attract the attention of researchers. Concatenative synthesis of Lithuanian speech was studied in the P. Kasparaitis papers (Kasparaitis, 1999; Kasparaitis, 2000; Kasparaitis, 2005). Methods of quality improvement in concatenative speech synthesis for the Polish language were considered in (Janicki, 2004). Formant Lithuanian vowel models have been developed in (Ringys and Slivinskas, 2009; Ringys and Slivinskas, 2010). A method for formant parameter extraction from a labelled single speaker database for use in a synthesis system is examined in (Mannell, 1998). A formant synthesis using rule-based and data-driven methods is presented in (R. Carlson et al., 2002).

The main concatenation synthesis problem is the size of the memory for storing the vocabulary. The synthesized speech quality does not achieve the natural speech quality since glitches occur on the concatenation boundaries.

Many synthesizers that use formant synthesis produce artificial speech that sound robot-like. Formant synthesizers, however, have advantages against the concatenative ones. The speech produced by a formant synthesizer can be sufficiently intelligible even at high speed. High speed speech synthesizing is necessary for screen reading programs. Additional advantages of formant synthesizers against the concatenative ones are the following: formant synthesizers require less computer memory than concatenative ones as they need no speech unit database. Formant synthesizers can control prosody aspects of the synthesized speech (intonation, rhythm, stress).

The most known Lithuanian speech synthesizer is based on concatenative synthesis (Balbonas, 2009). The practical implementation can be seen in (Garsiai.lt).

2.3. The signal fundamental frequency

Estimation of a fundamental frequency is very important in many fields of speech signal processing such as speech coding, speech synthesis, speech and speaker recognition (Cheveigné and Kawahara, 2002; Milivojevic et al., 2006). The speech signal fundamental frequency is an essential feature of human voice (Hess, 1983). The fundamental frequency is denoted by f_0. What we hear as a single sound when someone is speaking (for example, pronouncing /a/) is really the fundamental frequency plus a series of harmonics. The fundamental frequency is determined by the number of times the vocal folds vibrate in one second, and measured in cycles per second [cps], or Hertz [Hz]. The harmonics are multiples of the fundamental frequency. Thus if the fundamental frequency is 100 Hz, the harmonics are 200 Hz, 300 Hz, 400 Hz, etc. The fundamental frequency is also called the first harmonic. We normally don't hear the harmonics as separate tones, they, however, exist in the sound and add a lot of richness to the sound. Often the sinusoid of the frequency $f_k = kf_0$ is itself called the *k*-th harmonic of the signal.

In order to get good quality of a synthesised sound, one needs to estimate this frequency as accurately as possible. The discrete Fourier transform (DFT) method is usually used to estimate this frequency. This method gives good results when the observed signal is sufficiently long. For shorter signals, performance of this method is not satisfactory. Thus alternative methods have to be used. One of such algorithms is the so-called MUSIC method. This method is used widely in the mobile communications field. In (Murakami and Ishida, 2001), T. Murakami and Y. Ishida applied the MUSIC method for the

analysis of speech signals. They used this method for the fundamental frequency estimation of Japanese female and male vowels /a/, /e/, /i/, /o/, /u/, and illustrated that their method based on the MUSIC method is superior to the conventional cepstral method for estimating the fundamental frequency.

MUSIC method

Consider the following model:

$$y_n = \sum_{l=1}^{p} c_l \exp(jw_l n) + e_n, \quad (n = 1, \cdots, N) \tag{1}$$

where $c_l \in C$, $\{e_n\}$ is white noise. Let M be some integer greater than p.

Define:

$$\begin{aligned} &\mathbf{y}(t) = [y_t, \ldots, y_{t+M-1}]^T, \quad \mathbf{x}(t) = [c_1 e^{jw_1 t}, \ldots, c_p e^{jw_p t}]^T, \\ &\mathbf{e}(t) = [e_1, \ldots, e_{t+M-1}]^T \end{aligned} \tag{2}$$

where $t = 1, \ldots, N - M + 1$. Define also:

$$\begin{aligned} &\mathbf{a}(w) = [1, e^{jw}, \ldots, e^{j(M-1)w}]^T, \quad \boldsymbol{\theta} = [w_1, \ldots, w_p]^T, \\ &\mathbf{A}(\theta) = [a(w_1), \ldots, a(w_p)]. \end{aligned} \tag{3}$$

We can now write (1) as

$$\mathbf{y}(t) = \mathbf{A}(\theta)\mathbf{x}(t) + \mathbf{e}(t) \quad (t = 1, 2, \ldots, K = N - M + 1) \tag{4}$$

The MUSIC method (Schmidt, 1986; Stoica and Moses, 1997; Therrien, 1992) was developed in 1979 by American scientist R. Schmidt. The acronym MUSIC stands for MUltiple SIgnal Classification. This method deals with estimation of parameters of (4) model.

The covariance matrix $\boldsymbol{R} = E\boldsymbol{y}(t)\boldsymbol{y}^H(t)$ of the vector $\boldsymbol{y}(t)$ is given by

(Stoica and Nehorai, 1989)

$$\mathbf{R} = \mathbf{A}(\theta)\mathbf{P}\mathbf{A}^H(\theta) + \sigma^2\mathbf{I}_{M\times M} \tag{5}$$

where σ^2 is as in $E\boldsymbol{e}(t)\boldsymbol{e}^H(t) = \sigma^2\boldsymbol{I}_{M\times M}$, and $\boldsymbol{P} = E\boldsymbol{x}(t)\boldsymbol{x}^H(t)$.

Denote by $\gamma_1 > \gamma_2 > \cdots > \gamma_M$ the eigenvalues of the matrix $\boldsymbol{R}$. Since rank $(\mathbf{A}\mathbf{P}\mathbf{A}^H) = p$ (Stoica and Nehorai, 1989), then

$$\gamma_k > \sigma^2 \ (k = 1, \ldots, p) \ \text{ and } \ \gamma_k = \sigma^2 \ (k = p+1, \ldots, M). \tag{6}$$

Let $s_1, s_2, \ldots, s_p$ be the unit-norm eigenvectors corresponding to the first p largest eigenvalues $\gamma_1, \gamma_2, \ldots, \gamma_p$, and $g_1, g_2, \ldots, g_{M-p}$ - the unit-norm eigenvectors corresponding to the last *M-p* smallest eigenvalues $\gamma_{p+1}, \gamma_{p+2}, \ldots, \gamma_M$. Denote by $\boldsymbol{S}$ an $M\times p$ matrix whose columns are the vectors $s_1, s_2, \ldots, s_p$, and by $\boldsymbol{G}$ an $M\times(M-p)$ matrix whose columns are the vectors $g_1, g_2, \ldots, g_{M-p}$, i. e.

$$\mathbf{S} = [s_1, \ldots, s_p], \quad \mathbf{G} = [g_1, \ldots, g_{M-p}]. \tag{7}$$

It is shown in (Stoica and Nehorai, 1989) that the true parameter values $\{w_1, \ldots, w_p\}$ are the only solutions of the following equation: $\mathbf{a}^H(w)\mathbf{G}\mathbf{G}^H\mathbf{a}(w) = 0$.

In practice, we use an estimate

$$\hat{R} = \frac{1}{M}\sum_{t=1}^{M}\mathbf{y}(t)\mathbf{y}^H(t) \tag{8}$$

of the true covariance matrix $\boldsymbol{R}$. Denote by $\hat{s}_1, \ldots, \hat{s}_p, \hat{g}_1, \ldots, \hat{g}_{M-p}$ the unit-norm eigenvectors of $\widehat{\boldsymbol{R}}$ arranged in the descending order of the corresponding

eigenvalues, and by $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{G}}$ - the matrices made of $\{\hat{s}_1, \ldots, \hat{s}_p\}$ and $\{\hat{g}_1, \ldots, \hat{g}_{M-p}\}$. Define the MUSIC spectral function as follows:

$$\hat{\mathbf{P}}_{MU}(e^{jw}) = \frac{1}{\mathbf{a}^H(e^{jw})\hat{\mathbf{G}}\hat{\mathbf{G}}^H\mathbf{a}(e^{jw})}. \quad (9)$$

The estimates of $\{w_1, \ldots, w_p\}$ are obtained by maximizing $\hat{\mathbf{P}}_{MU}(e^{jw})$. This procedure is done by evaluating it at the points of a fine grid.

2.4. Diphone an phone concepts

A phone is an individual sound unit of speech without concern as to whether or not it is a phoneme of some language (Onelook 2010). Remind that a phoneme is any of the distinct units of sound that distinguishes one word from another, e.g. *p*, *b*, *d*, and *t* in *pad*, *pat*, *bad*, and *bat* (Oxford dictionaries 2010).

A phone can also be defined as an unanalysed sound of a language. It is the smallest identifiable unit found in a stream of speech that is able to be transcribed with an IPA symbol (SIL 2004) where IPA stands for the abbreviation International Phonetic Alphabet. The IPA is a system of phonetic notation based primarily on the Latin alphabet, devised by the International Phonetic Association as a standardized representation of the sounds of spoken language (International Phonetic Association, 1999). The IPA is designed to represent only those qualities of speech that are distinctive in spoken language: phonemes, intonation, and the separation of words and syllables.

A word „diphone“ can be derived from two Greek words „di“ that means „two“ and “phonos” that means “sound”. Diphones contain the transitions from one sound to the next and form building blocks for synthetic speech. Spanish has about 800 diphones, English – over 1500, German – about 2500, and Lithuanian – about 5000 (Cressey, 1978; Fox, 2005). By combining pre-

recorded diphones, we can create much more natural synthesized speech sounds than by combining just simple phones, because the pronunciation of each phone varies depending on the surrounding phones.

2.5. Lithuanian speech sounds

Lithuanian language phonemes (sounds) were studied in (Girdenis, 1995). In this work, A. Girdenis listed 58 phonemes. All these phonemes are unstressed. In order to take stress into account, this list was appended by 29 stressed phonemes (Kasparaitis, 2005). A study of Lithuanian compound diphthongs suggested including 4 additional phonemes (Kasparaitis, 2005). All the phonemes mentioned above along with a pause make a list of 92 units. P. Kasparaitis has developed a system of coding of Lithuanian phonemes. The Lithuanian phoneme list along with the examples is presented in Table 1 of the Appendix A.

2.5.1. Lithuanian vowel sounds

Table 1 of the Appendix A shows that Lithuanian language has twenty eight pure vowel phonemes. Five of them are marked with a letter "**a**", five - with a letter "**e**", three - with a letter "**ė**'', five - with a letter "**i**", five - with a letter "**o**", and five - with a letter "**u**". The plots of the vowel phoneme signals are presented in Fig. 1 – 6 of the Appendix B. These plots show that the periods of the stressed vowels are longer than those of the unstressed vowels, i. e. the fundamental frequencies – that are the period reciprocals – of the stressed vowels are lower than those of the unstressed ones. The fundamental frequency values of all unstressed and stressed vowels using 50 utterances by female have been measured. The averages values of the fundamental frequencies and their confidence intervals are shown in Table 1. The confidence intervals are stated at the 95 % confidence level.

Table 1 *The values of the fundamental frequencies of unstressed and stressed vowel phonemes (female speaker)*

No	Phoneme	Example	Frequency [Hz]	Confidence intervals
1	/a/	m**a**mà 'mother'	236	[234.5, 237.4]
2	/a`/	lazd**à** 'stick'	152	[150.5, 153.4]
3	/a:/	dr**ą**sà 'courage'	200	[198.3, 201.7]
4	/a:´/	k**á**rdas 'sword'	169	[167.0, 171.0]
5	/a:~/	**ã**čiū 'thank you'	162	[160.1, 163.8]
6	/e/	m**e**dãlis 'medal'	238	[236.3, 239.7]
7	/e`/	sug**è**sti 'turn bad'	156	[154.5, 157.5]
8	/e:/	gr**ę**žinỹs 'well'	270	[268.6, 271.4]
9	/e:´/	**é**rkė 'mite'	177	[175.7, 178.3]
10	/e:~/	gyv**ẽ**nimas 'life'	171	[169.8, 172.2]
11	/ė:/	k**ė**dẽ 'chair'	240	[238.7, 241.3]
12	/ė:´/	up**é**takis 'trout'	175	[173.8, 176.2]
13	/ė:~/	gėl**ẽ** 'flower'	162	[160.9, 163.1]
14	/i/	l**i**gà 'disease'	219	[217.7, 220.3]
15	/i`/	k**ì**škis 'rabbit'	185	[183.7, 186.3]
16	/i:/	t**ỹ**là'silence'	278	[276.8, 279.2]
17	/i:´/	r**ý**tas 'morning'	179	[177.7, 180.3]
18	/i:~/	arkl**ỹ**s 'horse'	174	[172.5, 175.5]
19	/o/	**o**žkà 'she-goat'	273	[271.6, 274.4]
20	/o`/	ch**ò**ras 'choir'	166	[164.6, 167.4]
21	/o:/	k**o**vótojas 'fighter'	228	[226.7, 229.3]
22	/o:´/	š**ó**nas 'side'	180	[178.6, 181.4]
23	/o:~/	Ad**õ**mas 'Adam'	183	[181.8, 184.2]
24	/u/	k**u**ltūrà 'culture'	231	[230.0, 232.0]
25	/u`/	**ù**pė 'river'	189	[187.8, 190.2]
26	/u:/	k**ū**rinỹs 'work'	203	[201.7, 204.3]
27	/u:´/	l**ú**pa 'lip'	183	[181.5, 184.5]
28	/u:~/	m**ũ**šis 'battle'	180	[178.6, 181.4]

The bar plots of the vowel phoneme fundamental frequency values are shown in Fig. 1.

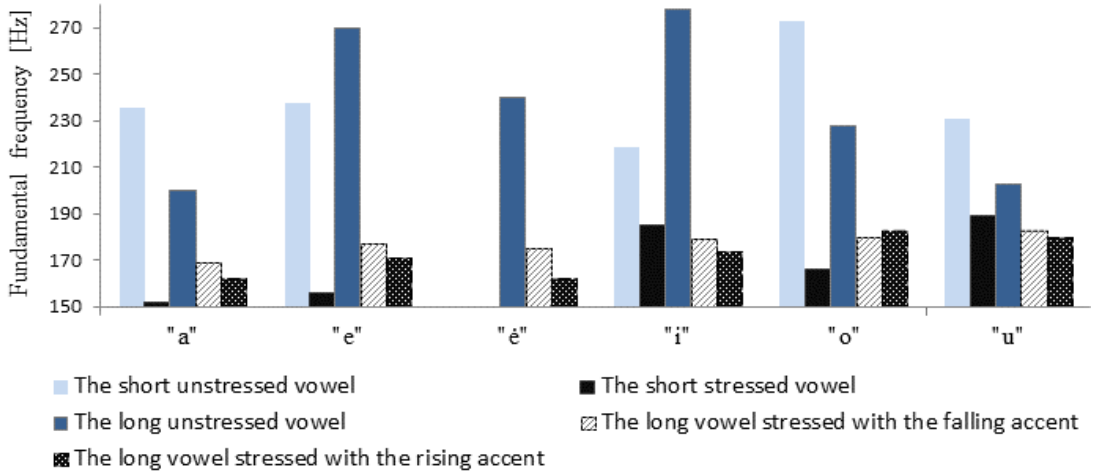


Fig. 1 *The tendencies of the vowel phoneme fundamental frequency changing*

This figure shows that the unstressed vowels have longer bars in comparison with the ones of the stressed vowels. Note that the vowel "ė" has no short phonemes therefore it has only three bars.

2.5.2. Lithuanian semivowel sounds

The Lithuanian consonants "j", "l", "m", "n", "r", "v" are called semivowels as they have both vowel and semivowel features. Lithuanian language has nineteen pure semivowel phonemes (see Table 1 of the Appendix A). One of them is marked with a letter "j", four - with a letter "l", four - with a letter "m", four - with a letter "n", four - with a letter "r", and two - with a letter "v". The plots of the semivowel phoneme signals are presented in Fig. 7 – 12 of the Appendix C. These plots show that the periods of the stressed semivowels are longer than those of the unstressed semivowels, i. e. the fundamental frequencies – that are the period reciprocals – of the stressed semivowels are lower than those of the unstressed ones. The fundamental frequency values of all unstressed and stressed semivowels using 50 utterances by female have been measured. The averages values of the fundamental frequencies and their confidence intervals are shown in Table 2. The confidence intervals are stated at the 95 % confidence level.

Table 2 *The values of the fundamental frequencies of unstressed and stressed semivowel phonemes (female speaker)*

No	Phoneme	Example	Frequency [Hz]	Confidence intervals
1	/j"/	**j**ūra 'sea'	170	[168.6, 171.4]
2	/l/	vá**l**sas 'waltz'	167	[165.5, 168.5]
3	/l˜/	vi**l̃**kas 'wolf'	156	[154.6, 157.4]
4	/l"/	va**l**ià 'will'	183	[181.4, 184.5]
5	/l"˜/	gu**l̃**ti 'to go to bed'	160	[158.4, 161.6]
6	/m/	ã**m**atas 'handicraft'	173	[171.7, 174.3]
7	/m˜/	li**m̃**palas 'adhesive'	167	[165.8, 168.2]
8	/m"/	s**m**ẽgenys 'brain'	182	[180.8, 183.2]
9	/m"˜/	ka**m̃**štis 'cork'	169	[167.7, 170.3]
10	/n/	**n**ãmas 'house'	215	[213.5, 216.5]
11	/n˜/	i**ñ**karas 'anchor'	185	[183.4, 186.6]
12	/n"/	**n**ėšti 'to carry along'	202	[200.6, 203.4]
13	/n"˜/	le**ñ**ktis 'to bend'	181	[179.6, 182.4]
14	/r/	**r**ãtas 'wheel'	210	[208.4, 211.6]
15	/r˜/	ga**r̃**sas 'sound'	178	[176.6, 179.5]
16	/r"/	k**r**iáušė 'pear'	219	[217.6, 220.4]
17	/r"˜/	ki**r̃**tis 'stress', 'blow'	182	[180.5, 183.5]
18	/v/	**v**óras 'spider'	185	[183.4, 186.6]
19	/v"/	**v**iáuksėti 'to yelp'	156	[154.6, 157.4]

The bar plots of the semivowel phoneme fundamental frequency values are shown in Fig. 2.

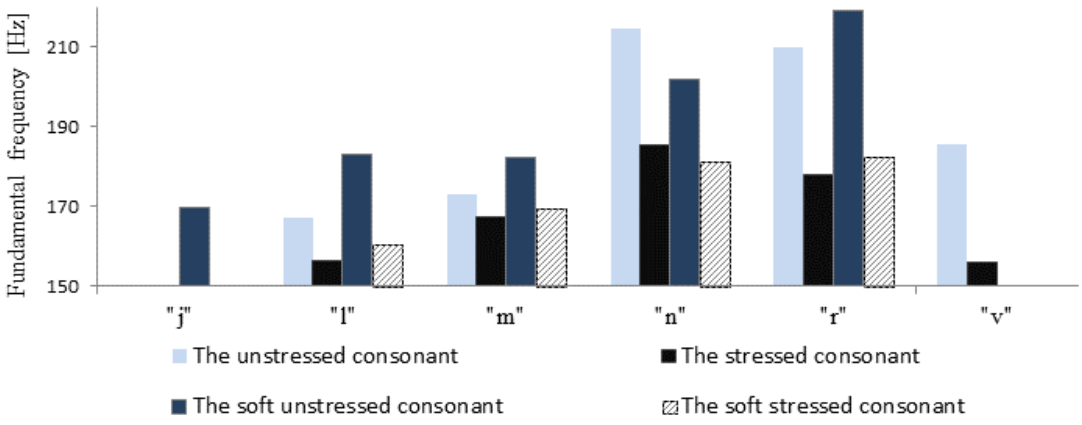


Fig. 2 *The tendencies of the semivowel phoneme fundamental frequency changing*

Fig. 2 shows that the unstressed vowels have longer bars in comparison with the ones of the stressed vowels.

2.5.3. Lithuanian diphthong sounds

A diphthong is defined as a complex speech sound or glide that begins with one vowel and gradually changes to another vowel within the same syllable, as (oi) in boil or ($\bar{i}$) in fine (Collins, 2009).

Lithuanian language has 6 pure diphthongs and 16 mixed diphthongs. The pure diphthongs consist of two vowels and are the following: ai, au, ei, ie, ui, uo. A mixed diphthong is a complex speech sound that begins with a short vowel (*i, e, u, a*) and ends with a consonant (*l, r, m, n*) within the same syllable (Garšva, 2001).

There are two types of pure diphthongs: gliding diphthongs (ie, uo) and compound diphthongs (ai, au, ei, ui). We do not feel a phonetic boundary between the first and the second element in pronouncing the gliding diphthongs (LRC, 2013). The gliding diphthongs are sometimes called the complex diphthongs (Garšva, 2001).

A list of compound diphthongs and corresponding phonemes or phoneme combinations is presented in Table 3. Table 4 contains a list of gliding diphthongs and corresponding phonemes or phoneme combinations. The phoneme notation presented in Table 3 and Table 4 corresponds to the notation used in (Kasparaitis, 2005).

Table 3 *The compound diphthongs and corresponding phoneme combinations*

ai		**au**	
láimė 'happiness'	/Aa/+/j/	káukė 'mask'	/Aa/+/w/
laĩkas 'time'	/a/+/J/	baũbas 'bugaboo'	/a/+/W/
vaikaĩ 'children'	/a/+/j/	laukaĩ 'fields'	/a/+/w/
ei		**ui**	
éižėti 'to crack'	/Ea/+/j/	užgùiti 'to oppress'	/U/+/j/
peĩlis 'knife'	/e/+/J/	zuĩkis 'rabbit'	/u/+/J/
eiklùs 'nimble'	/e/+/j/	puikùs 'excellent'	/u/+/j/

Table 4 *The gliding diphthongs and corresponding phonemes*

ie		**uo**	
síela 'soul'	/Ie/	púodas 'pot'	/Uo/
kriẽnas ' horseradish'	/iE/	pietuõs 'in the south '	/uO/
kiemaĩ 'yards'	/ie/	puodẽlis 'cup'	/uo/

In Lithuanian speech, a pure diphthong can be stressed or unstressed. If the first vowel of the diphthong is stressed, then the syllable has the falling (acute ‚/‘) accent. If the second vowel of the diphthong is stressed, then the syllable has the rising (circumflex ‚~‘) accent.

Three other diphthongs (eu, oi, ou) can be observed only in international words, e.g., eukaliptas 'eucalyptus', sinusoidė 'sinusoid', šou 'show'.

Table 5 *Lithuanian diphthongs met in international words only and the corresponding phoneme combinations*

Eu		**Oi**		**ou**	
terapèutas 'physician'	/E/+/u/	bòileris 'boiler'	/O/+/i/	klòunas 'clown'	/O/+/u/
				noùmenas 'noumenon'[1]	/o/+/U/
Europà 'Europe'	/e/+/u/	boikòtas 'boycott'	/o/+/i/	drèdnoutas 'dreadnought'[2]	/o/+/u/

The number of vowels and compound/complex diphthongs or mixed

[1] noumenon – (in the philosophy of Immanuel Kant) a thing as it is in itself, not perceived or interpreted, incapable of being known, but only inferred from the nature of experience (Collins, 2009). The term is used in contrast with, or in relation to what Kant called the phenomenon – the thing as it appears to an observer (Encyclopedia Britannica, 2012).

[2] dreadnought – 1) (*historical*) a type of battleship introduced in the early 20th century, larger and faster than its predecessors and equipped entirely with large-calibre guns; 2) (*archaic*) a fearless person; 3) (*archaic*) a heavy overcoat for stormy weather (Oxford, 2010).

diphthongs defines the number of syllables in a word. Compound/complex diphthongs are important for Lithuanian language as they along with vowels and mixed diphthongs make the basis of a Lithuanian language syllable. Diphthong modelling is important for solving a text-to-speech (TTS) problem.

One can find research papers that investigate diphthongs of various languages in the literature. Acoustic analysis of the Spanish diphthongs was carried out in (Borzone De Manrique, 1979). The German diphthongs were analysed in (Geumann, 1997). Acoustic analysis of diphthongs in Standard South African English was carried out in (Martirosian and Davel, 2008).

2.6. Conclusions of Section 2

The dominating fields of Lithuanian speech engineering are speech recognition and speech synthesis.

There exist two main speech synthesis methodologies: concatenation synthesis and formant synthesis. Concatenation synthesis relies on a speech sound recorded in advanced database, meanwhile formant synthesis – on mathematical sound models.

Lithuanian language has ninety two phonemes. Twenty eight of them are pure vowel phonemes, nineteen – semivowel phonemes.

The differences between the fundamental frequency values of unstressed and stressed vowel and semivowel phonemes have been ascertained. Investigations have shown that the fundamental frequencies of the stressed vowels and semivowels are lower than those of the unstressed ones.

Lithuanian language has six pure diphthongs and sixteen mixed diphthongs. The pure diphthong consist of two vowels, a mixed diphthong is a complex speech sound that begins with a short vowel and ends with a consonant within the same syllable.

3

Phoneme modelling framework

Lithuanian vowel and semivowel phoneme modelling framework based on a vowel and semivowel phoneme mathematical model and an automatic procedure of estimation of the vowel phoneme fundamental frequency and input determining is proposed. Using this framework, the phoneme signal is described as the output of a linear multiple-input and single-output (MISO) system. The MISO system is a parallel connection of single-input and single-output (SISO) systems whose input impulse amplitudes vary in time. Within this framework two synthesis methods are proposed: harmonic and formant.

3.1. Vowel and semivowel phoneme signals decomposition into harmonics

The goal is to get mathematical models of the analysed phoneme, which could be used as a base of a phoneme synthesizer. In general case, the character of vowel and semivowel signals is periodic (see Fig. 3).

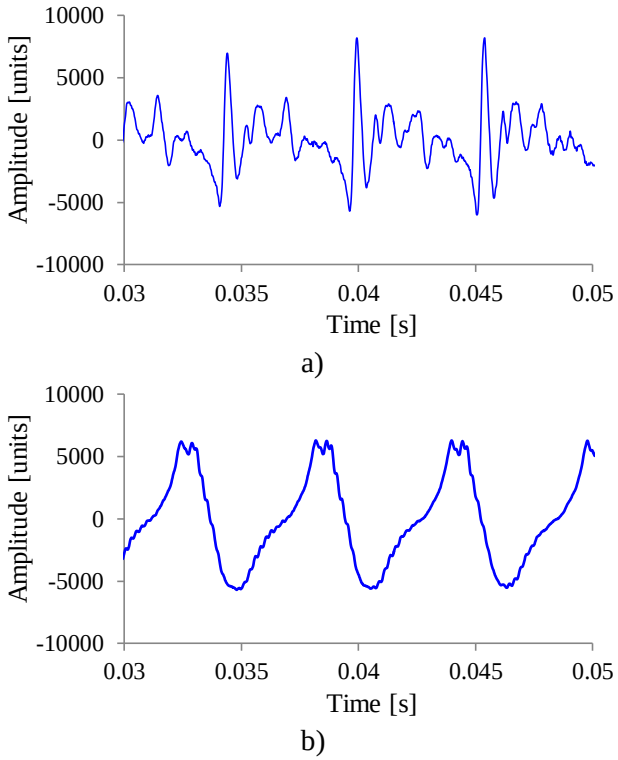


Fig. 3 *The periodic character of phonemes: a) the plot of the vowel /a/, b) the plot of the semivowel /m/*

One can see from Fig. 3 that a phoneme signal has a rather complex form. It is very difficult to find such a model that fits the phoneme signal well. The approach of expanding a complex signal into the sum of simpler signals is used. The signal can be expanded into the sum of a finite number of components – the phoneme signal harmonics (similarly as harmonics of a periodic signal in Fourier series theory).

Suppose the phoneme signal $s(n)$ can be expanded into the sum of L harmonics:

$$s(n) = s_1(n) + s_2(n) + \cdots + s_L(n), \quad n = 1, \dots, N \tag{10}$$

where L is the number of harmonics, N is the number of samples of the phoneme.

In order to decompose the phoneme signal into harmonics, the fundamental frequency of this signal has been estimated. The recurrent algorithm of

estimating the fundamental frequency and calculating harmonics is present below. At first, the magnitude response of the whole vowel phoneme signal is calculated. Let f_0 be an initial estimate of the fundamental frequency. Then the frequency band 0 – 6000 Hz is partitioned into the subbands shown in Table 6.

Table 6 *The frequency band partition into subbands*

Subband number	Subband
1	$[0.5f_0, 1.5f_0)$
2	$[1.5f_0, 2.5\ f_0)$
…	
L	$[\big((L-1)+0.5\big)f_0, (L+0.5)f_0)$

L is the largest integer number for which the inequality $(L+0.5)f_0 \leq 6000$ holds, i. e.

$$L = [6000/f_0 - 0.5] \tag{11}$$

where marking [ ] stands for the integer part of a real number. Note that we do not consider the records obtained with the sampling frequency lower than 12 000 Hz. For the higher values of the sampling frequency we consider the frequency band [0, 6000] Hz only. An example of a magnitude response and its partition into subbands is given in Fig. 4.

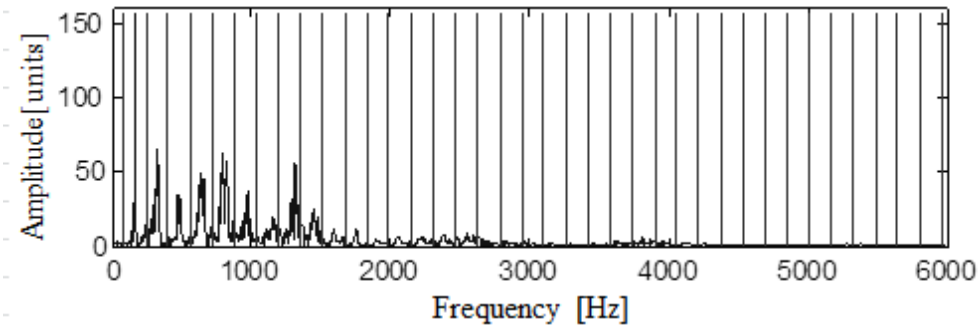


Fig. 4 *The magnitude response of a vowel and its partition into subbands*

In each subband, the highest amplitudes $a_1, a_2, \dots, a_L$, are determined, and the frequencies corresponding to those amplitudes: $g_1, g_2, \dots, g_L$ are found. At first glance, these frequencies look like the formants, this, however, is not true in general case. Then we compare the frequency sequences $f_0, 2f_0, \dots, Lf_0$ and $g_1, g_2, \dots, g_L$. Our goal is to find such an f_0 that minimizes the sum of the distances between the frequencies:

$$S_0 = \sum_{l=1}^{L} |lf_0 - g_k|. \quad (12)$$

The algorithm that achieves this goal is described below. The data of the algorithm is as follows:

1. The initial value of the fundamental frequency f_0.
2. The number of subbands (harmonics) L (L is defined by (11))
3. The values of the harmonic frequencies $g_1, g_2, \dots, g_L$.

The steps of the algorithm are listed below.

Step 1. Compute the sum of the distances $S_0 = \sum_{l=1}^{L} |lf_0 - g_l|$.

Step 2. Set $\Delta = 1$

(Δ is the difference between the new fundamental frequency value f_{new} and the old fundamental frequency value f_0).

Step 3. Compute the new fundamental frequency value $f_{new} = f_0 + \Delta$.

Step 4. Compute the sum of the distances $S_{new} = \sum_{l=1}^{L} |lf_{new} - g_k|$.

Step 5. If $S_{new} < S_0$ then

$f_0 = f_{new}, S_0 = S_{new}$

else

$f_{new} = f_0 - \Delta, S_{new} = \sum_{l=1}^{L} |lf_{new} - g_l|$

if $S_{new} < S_0$ then

$f_0 = f_{new}, S_0 = S_{new}$

else

$\Delta = \Delta/2$.

Step 6. If $\Delta < 0.01$ then

Go to Step 7

else

Go to Step 3

Step 7. END

We denote the obtained value by $\widetilde{f_0}$.

A block diagram of the algorithm presented above is shown in Fig. 5.

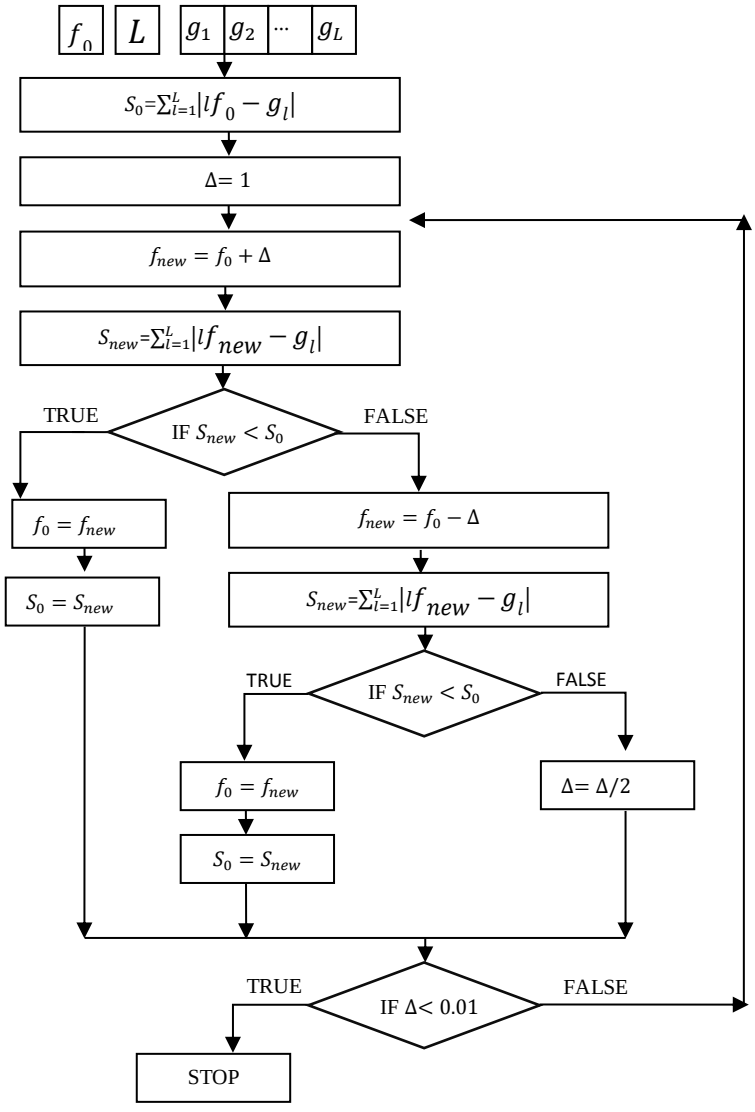


Fig. 5 *A block diagram of the fundamental frequency refining algorithm*

At first glance it may seem that the cycle becomes infinite if the condition $S_{new} < S_0$ is always true. In practice, when the f_0 changes, the distance between the values lf_0 and g_l can not decrease all the time; at a certain time it will start to increase.

After obtaining the optimal value $\widetilde{f_0}$, we can decompose the phoneme signal $s(n)$ into L harmonics (L is defined by (11)). For this purpose, a new frequency band partition into subbands according to Table 6 (with $f_0 = \widetilde{f_0}$) is made.

An auxiliary function $g_l(m)$ is defined as follows:

$$g_l(m) = \begin{cases} FFT(s(m)), m \in \left[\left((l-1)+0.5\right)\widetilde{f_0}, (l+0.5)\widetilde{f_0}\right] \\ 0, m \notin \left[\left((l-1)+0.5\right)\widetilde{f_0}, (l+0.5)\widetilde{f_0}\right], \end{cases} \quad (13)$$

where $l = 1, \dots, L$, FFT – the fast Fourier transform, and compute its inverse Fourier transform

$$\tilde{s}_l(n) = \left(\frac{1}{N}\right) \sum_{m=1}^{N} g_l(m) e^{(2\pi i)(n-1)\frac{m-1}{N}}. \quad (14)$$

$n = 1, \dots, N$, i – imaginary unit.

The obtained signal $\widetilde{s_l}$ is the l-th harmonic of the phoneme signal. The first three harmonics of the female vowel /a:˜/ are shown in Fig. 6.

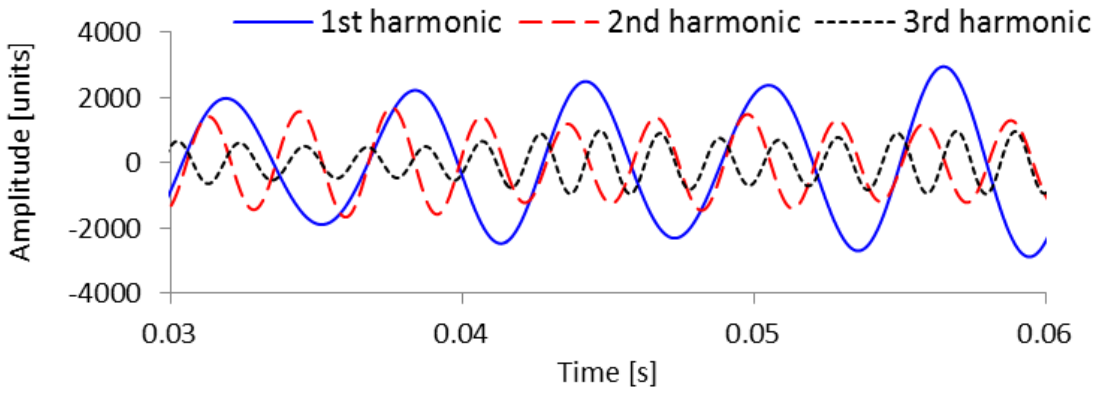


Fig. 6 *The first three harmonics of the phoneme /a:˜/ (as in the word ãčiū)*

We see from Fig. 6 that the harmonic amplitudes are not constant. Note that the harmonic periods are not constant, too. This changing over time of the amplitudes and periods gives sounding naturalness.

3.2. Vowel and semivowel phoneme signals decomposition into formants

A formant (formant frequency) is defined in a usual way as a maximum of the phoneme spectrum envelope. In time-domain representation, a formant can be described as the output signal of the filter whose impulse response is a damped sinusoid (Fant, 1970; Cook, 2002). In the current work, the Linear Predictive Coding (LPC) method (Markel and Gray, 1976) as a formant extraction tool is used. We need to partition the frequency band 0-6000 Hz into the subbands where each band corresponds to one formant. The partition is executed in the following way. Using the LPC method, the signal envelope is obtained. The frequency values corresponding to the envelope local minima are considered as partition points. It is very important that the formant extraction coincide with the harmonic extraction, i. e. a part of a harmonic cannot belong to one formant and the other part belong to the other formant. Also, each harmonic should be assigned to a certain formant. I propose to add the neighbouring harmonics (calculated in Section 3.1) corresponding to a selected formant of spectrum. The obtained signal will be call a *formant*. Joining of the harmonic frequencies of the phoneme spectrum into groups corresponding to the particular formant frequencies is shown in Fig. 7.

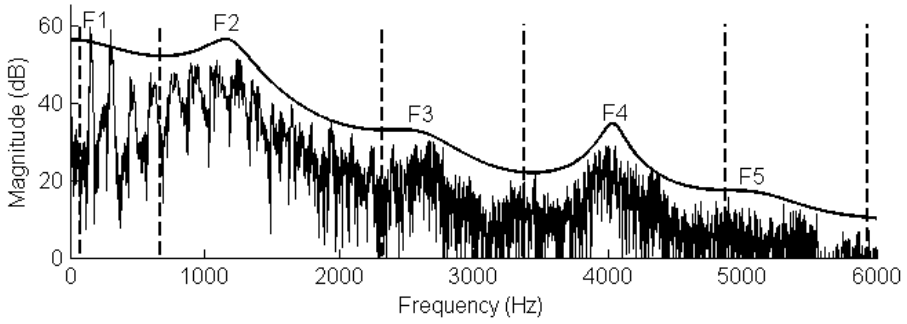


Fig. 7 *The plots of the spectrum of the phoneme /a:˜/*

The first three formant components of the female vowel /a:˜/ are shown in Fig. 8.

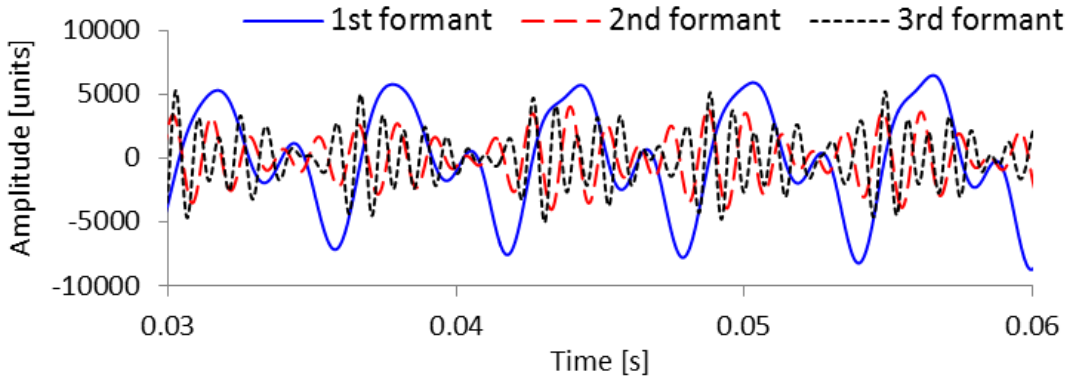


Fig. 8 *The first three formant components of the phoneme /a:˜/*

We see from Fig. 4 and Fig. 6 that in general case the form of a formant signal is more complex than that of a harmonic signal. Harmonics are more similar to sine waves, and the formants look like pulsating vibrations.

3.3. The vowel and semivowel phoneme model

We have to expand a phoneme signal into components. It is natural to choose a parallel connection model where each component is modelled separately. For modelling a discrete time linear stationary system with K inputs and a single output (K is the number of harmonics (in harmonic synthesis method) or formants (in formant synthesis method)) is used. The system is stationary as its

parameters are lumped, i. e. they do not vary in time for a selected phoneme. The system is linear since the output is a linear combination of the present and past values of the input signals.

A diagram of such a system is shown in Fig. 9.

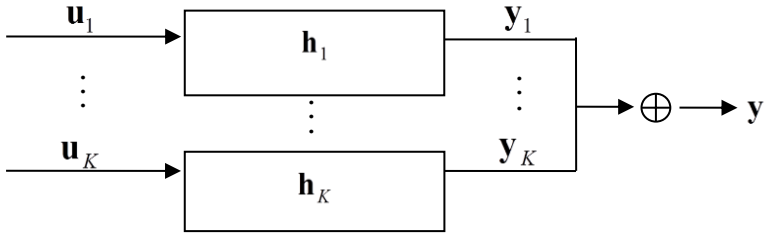


Fig. 9 *A MISO system for vowel and semivowel phonemes modelling*

Let

$$\begin{aligned} \mathbf{h}_1 &= (h_1(0),\, h_1(1),\, h_1(2),\, ...) \\ &\vdots \\ \mathbf{h}_K &= (h_K(0),\, h_K(1),\, h_K(2),\, ...) \end{aligned} \tag{15}$$

denote the impulse responses, and

$$\begin{aligned} \mathbf{u}_1 &= (\ldots,\, u_1(-1),\, u_1(0),\, u_1(1),\, ...) \\ &\vdots \\ \mathbf{u}_K &= (\ldots,\, u_K(-1),\, u_K(0),\, u_K(1),\, ...) \end{aligned} \tag{16}$$

– the inputs of the corresponding SISO systems. Denote by

$$\mathbf{y}_k = (\ldots,\, y_k(-1),\, y_k(0),\, y_k(1),\, ...)\,, \quad k = 1, \ldots, K \tag{17}$$

the output of the *k*-th SISO system.

The impulse response **H** of the MISO system is the following vector of sequences:

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_K \end{bmatrix}, \tag{18}$$

and the input of the MISO system is as follows: $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]^T$.

Such a MISO system whose output **y** is equal to the sum of the inputs of the SISO systems is considered:

$$\mathbf{y} = \mathbf{y}_1 + \ldots + \mathbf{y}_K . \tag{19}$$

Since

$$\boldsymbol{y}_k = \boldsymbol{u}_k * \boldsymbol{h}_k \tag{20}$$

where $*$ denotes the convolution operation

$$\mathbf{u}_k * \mathbf{h}_k = \sum_{i=0}^{\infty} u_k(n-i) h_k(i) , \tag{21}$$

we have

$$\boldsymbol{y} = \boldsymbol{u}_1 * \boldsymbol{h}_1 + \cdots + \boldsymbol{u}_K * \boldsymbol{h}_K . \tag{22}$$

Consider the vector delta sequence $\boldsymbol{\delta}(n)$ defined as

$$\boldsymbol{\delta}(n) = \begin{cases} [1,\ 1,\ \ldots,\ 1]^T, & n = 0 \\ [\underbrace{0,\ 0,\ \ldots,\ 0}_{K}]^T, & n \neq 0 \end{cases} . \tag{23}$$

Suppose that the system is excited by this sequence. In this case, the output sequence values are as follows:

$$\mathbf{y}(n) = \sum_{k=1}^{K} \mathbf{h}_k(n), \quad n \geq 0 . \tag{24}$$

The chosen model allows exciting each channel with a separate input sequence. This enables us to preserve harmonic amplitude variation.

If we take a single period of the phoneme component we get a signal of a certain form. This signal is similar to a quasipolynomial that is the product of a sinusoid and polynomial. It is assumed that all the $\mathbf{h}_k$ can be represented by second degree quasipolynomials. The mathematical description of a second degree continuous-time quasipolynomial is given below:

$$q(t) = e^{\lambda t}(a_1 \sin(2\pi f t + \varphi_1) + a_2 t \sin(2\pi f t + \varphi_2) + a_3 t^2 \sin(2\pi f t + \varphi_3)) \quad (25)$$

where t is a nonnegative real number $t \in R^+ \cup \{0\}$, $\lambda < 0$ - the damping factor, f – the frequency, a_k - amplitude, φ_k $(-\pi \le \varphi_k < \pi)$ – phase. If the coefficients $a_2 = a_3 = 0$, then we obtain the usual mathematical description of a formant.

First, consider a discrete time SISO system whose impulse response $\mathrm{h}_k(n)$ consists of a single component:

$$h_k(n) = q(n\Delta t), \quad n = 0, 1, 2, \ldots \quad (26)$$

where the component is defined by (25) with $\lambda = \lambda_k$, $f = f_k$, $a_1 = a_{k1}$, $a_2 = a_{k2}$, $a_3 = a_{k3}$, $\varphi_1 = \varphi_{k1}$, $\varphi_2 = \varphi_{k2}$, $\varphi_3 = \varphi_{k3}$, $\Delta t = 1/f_s$ (f_s is the sampling frequency), $k = 1, \ldots, K$.

Let N be a positive integer. Taking $t = 0, \Delta t, 2\Delta t, \ldots, (N-1)\Delta t$, we get from (25) and (26) that

$$
\begin{aligned}
&\mathrm{h}_k(0) = a_{k1}\sin(\varphi_{k1}) \\
&\mathrm{h}_k(\Delta t) = e^{\Delta t\lambda_k}(a_{k1}\sin(2\pi f_k\Delta t + \varphi_{k1}) + a_{k2}\Delta t\sin(2\pi f_k\Delta t + \varphi_{k2}) + a_{k3}\Delta t^2\sin(2\pi f_k\Delta t + \varphi_{k3})) \\
&\mathrm{h}_k(2\Delta t) = e^{2\Delta t\lambda_k}(a_{k1}\sin(2\pi f_k 2\Delta t + \varphi_{k1}) + a_{k2}2\Delta t\sin(2\pi f_k 2\Delta t + \varphi_{k2}) + a_{k3}2^2\Delta t^2\sin(2\pi f_k 2\Delta t + \varphi_{k3})) \\
&\vdots \\
&\mathrm{h}_k((N-1)\Delta t) = e^{(N-1)\Delta t\lambda_k}(a_{k1}\sin(2\pi f_k(N-1)\Delta t + \varphi_{k1}) + a_{k2}(N-1)\Delta t\sin(2\pi f_k(N-1)\Delta t + \varphi_{k2}) + \\
&\qquad a_{k3}(N-1)^2\Delta t^2\sin(2\pi f_k(N-1)\Delta t + \varphi_{k3})).
\end{aligned}
\tag{27}
$$

Denote by $\mathbf{h}_k^N$ the following $N \times 1$ vector :

$$
\mathbf{h}_k^N = [h_k(0), \quad h_k(\Delta t), \quad h_k(2\Delta t), \quad \ldots, \quad h_k((N-1)\Delta t)]^T . \tag{28}
$$

Let

$$
\begin{aligned}
&\Omega_k = 2\pi f_k\Delta t, \\
&\Lambda_k = \lambda_k\Delta t, \\
&A_{k1} = a_{k1}, \\
&A_{k2} = a_{k2}\Delta t, \\
&A_{k3} = a_{k3}\Delta t^2, \quad k = 1, ..., K.
\end{aligned}
\tag{29}
$$

Collect the first N values of the output sequence into a vector $\mathbf{y}^N$:

$$
\mathbf{y}^N = \left[y(0), \quad y(1), \quad y(2), \quad \ldots, \quad y(N-1)\right]^T . \tag{30}
$$

Define by $\mathbf{\Psi}_k = \mathbf{\Psi}_k(\Lambda_k, \Omega_k)$ the following $N \times 6$ matrix:

$$
\mathbf{\Psi}_k = [\mathbf{\Psi}_{k1} \quad \mathbf{\Psi}_{k2} \quad \mathbf{\Psi}_{k3}] \tag{31}
$$

where

$$
\mathbf{\Psi}_{ki} = \begin{bmatrix} \delta(i-1) & 0 \\ c_{ki} & s_{ki} \end{bmatrix} \qquad \left(\delta(i) = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0 \end{cases} \right) \tag{32}
$$

with

$$c_{ki} = \left[e^{\Lambda_k} \cos\Omega_k, \quad 2^{i-1} e^{2\Lambda_k} \cos 2\Omega_k, \quad \ldots, \quad (N-1)^{i-1} e^{(N-1)\Lambda_k} \cos(N-1)\Omega_k \right]^T \tag{33}$$

$$s_{ki} = \left[e^{\Lambda_k} \sin\Omega_k, \quad 2^{i-1} e^{2\Lambda_k} \sin 2\Omega_k, \quad \ldots, \quad (N-1)^{i-1} e^{(N-1)\Lambda_k} \sin(N-1)\Omega_k \right]^T \tag{34}$$

and by $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}(A_{k1}, A_{k2}, A_{k3}, \varphi_{k1}, \varphi_{k2}, \varphi_{k3})$ the following 6×1 vector:

$$\boldsymbol{\alpha}_k = [A_{k1} \sin(\varphi_{k1}), \quad A_{k1} \cos(\varphi_{k1}), \quad A_{k2} \sin(\varphi_{k2}), \quad A_{k2} \cos(\varphi_{k2}), \quad A_{k3} \sin(\varphi_{k3}), \quad A_{k3} \cos(\varphi_{k3})]^T. \tag{35}$$

Collect all the matrices $\{\boldsymbol{\Psi}_k\}$ into an $N \times 6K$ matrix $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1 | \boldsymbol{\Psi}_2 | \ldots | \boldsymbol{\Psi}_K], \tag{36}$$

and all the vectors $\{\boldsymbol{\alpha}_k\}$ into a $6K \times 1$ vector $\mathbb{\alpha}$:

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T | \boldsymbol{\alpha}_2^T | \ldots | \boldsymbol{\alpha}_K^T]^T. \tag{37}$$

One can readily check that the vector $\mathbf{y}^N$ can be expressed as the product of the matrix $\boldsymbol{\Psi}$ and the vector $\boldsymbol{\alpha}$

$$\mathbf{y}^N = \boldsymbol{\Psi}_1 \boldsymbol{\alpha}_1 + \ldots + \boldsymbol{\Psi}_K \boldsymbol{\alpha}_K = \boldsymbol{\Psi} \cdot \boldsymbol{\alpha}. \tag{38}$$

If we have the matrix $\boldsymbol{\Psi}$, obtaining the $\boldsymbol{\alpha}$ reduces to a simple least squares fit:

$$\boldsymbol{\alpha} = \boldsymbol{\Psi}^+ \cdot \mathbf{y}^N \tag{39}$$

where $\boldsymbol{\Psi}^+$ denotes the pseudo-inverse of the matrix $\boldsymbol{\Psi}$, i. e. a $6K \times N$ matrix $\boldsymbol{\Psi}^+ = (\boldsymbol{\Psi}^T \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}$.

Let z denote the common shift operator. Then from (18) we can write

$$z\mathbf{H} = \begin{bmatrix} z\mathbf{h}_1 \\ \vdots \\ z\mathbf{h}_K \end{bmatrix} \tag{40}$$

where

$$z\mathbf{h}_k = (h_k(1), h_k(2), h_k(3), ...), \quad k = 1, ..., K \tag{41}$$

is the shifted sequence.

Suppose now that the system is excited by a periodic sequence of the vector delta sequences:

$$\boldsymbol{\delta}(n), \boldsymbol{\delta}(n-M), \boldsymbol{\delta}(n-2M), ... \tag{42}$$

where $0 < M < N$ is the period of this sequence such that

$$z^k\mathbf{H} = \left.\begin{bmatrix} 0, 0, 0, ... \\ \vdots \\ 0, 0, 0, ... \end{bmatrix}\right\} K \text{ rows} \tag{43}$$

for all $k \geq 3M$. One can readily check that the output sequence then satisfies the following relationships:

$$\begin{aligned} &\mathrm{y}(n) = 0 \text{ for } n < 0 \\ &\mathrm{y}(n) = \sum_{k=1}^{K} \mathrm{h}_k(n) \text{ for } 0 \leq n < M \\ &\mathrm{y}(n) = \sum_{k=1}^{K} (\mathrm{h}_k(n) + \mathrm{h}_k(n-M)) \text{ for } M \leq n < 2M \\ &\mathrm{y}(n) = \sum_{k=1}^{K} (\mathrm{h}_k(n) + \mathrm{h}_k(n-M) + \mathrm{h}_k(n-2M)) \text{ for} \\ &l \cdot M \leq n < (l+1) \cdot M, \quad l = 2, 3, ... \ . \end{aligned} \tag{44}$$

Let $\boldsymbol{\theta} = [\Lambda_1, \ \Omega_1, \ \ldots, \ \Lambda_K, \ \Omega_K]$. Introduce the following $M \times 6K$ matrix

$\mathbf{\Phi} = \mathbf{\Phi}(\boldsymbol{\theta})$ (I call it the convoluted matrix):

$$\mathbf{\Phi} = \mathbf{\Psi}(1:M,:) + \mathbf{\Psi}(M+1:2M,:) + \mathbf{\Psi}(2M+1:3M,:) \qquad (45)$$

where $\mathbf{\Psi}(m:n,:)$ denotes a submatrix of the matrix $\mathbf{\Psi}$ consisting of all the columns and the rows starting from the m-th row and ending with the n-th row. Note that the matrix $\mathbf{\Phi}$ depends on the damping factors Λ_k and the angular frequencies Ω_k (and does not depend on the amplitudes A_{k1}, A_{k2}, A_{k3} and phases $\varphi_{k1}, \varphi_{k2}, \varphi_{k3}$).

Collect all the M values of the first period of the periodic output sequence into a vector $\mathbf{y}^M$:

$$\mathbf{y}^M = \left[y(0),\quad y(1),\quad y(2),\quad \ldots,\quad y(M-1)\right]^T. \qquad (46)$$

Using (44), it is not difficult to check that

$$\mathbf{y}^M = \mathbf{\Phi} \cdot \boldsymbol{\alpha}. \qquad (47)$$

Usually we measure the output $y(t)$ with errors. Therefore an error component into (47) must be incorporated. This error is modelled by an additive white Gaussian noise. Collect all the noise values into a vector **e**:

$$\mathbf{e} = \left[e(0),\quad e(1),\quad e(2),\quad \ldots,\quad e(M-1)\right]^T. \qquad (48)$$

Then the vector **y** can be written in the following form:

$$\mathbf{y}^M = \mathbf{\Phi} \cdot \boldsymbol{\alpha} + \mathbf{e}. \qquad (49)$$

We have to minimize the following functional:

$$r(\mathbf{a}, \boldsymbol{\theta}) = \left\| \mathbf{y}^M - \mathbf{\Phi}(\boldsymbol{\theta})\boldsymbol{\alpha} \right\|^2 \qquad (50)$$

where $\|\cdot\|$ is the Euclidian norm defined as $\|\mathbf{x}\|^2 = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{x}$.

It is proved in the literature (Golub and Pereyra, 1973) that minimization of the $r(\mathbf{a}, \boldsymbol{\theta})$ is equivalent to minimization of the following functional:

$$r_2(\boldsymbol{\theta}) = \left\| P^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})} y^M \right\|^2 \tag{51}$$

where

$$P^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})} = \mathrm{I}_M - \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}} = I_M - \boldsymbol{\Phi}\boldsymbol{\Phi}^{+}. \tag{52}$$

$P^{\perp}_{\Phi(\theta)}$ is called the orthogonal projector onto the orthogonal complement of the matrix $\boldsymbol{\Phi}$ column space. $P^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}$ can be represented by an $M \times M$ matrix. Equivalence of minimization can be explained as follows: suppose we found the value $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$ that minimizes $r_2(\boldsymbol{\theta})$. Then the value $(\hat{\mathbf{a}}, \hat{\boldsymbol{\theta}})$ where

$$\hat{\mathbf{a}} = \boldsymbol{\Phi}^{+}(\hat{\boldsymbol{\theta}})\mathbf{y}^M \tag{53}$$

minimizes the functional $r(\mathbf{a}, \boldsymbol{\theta})$. Levenberg-Marquardt approach to minimize the functional (51) is used. In the next Section, a new algorithm applied to the convoluted data is developed. The algorithm for nonconvoluted data was developed in (Šimonytė and Slivinskas, 1997).

3.4. Parameter estimation of the model

Levenberg-Marquardt approach (Levenberg, 1944; Marquardt, 1963) is an iterative procedure that corrects an initial parameter estimate according to the following formula:

$$\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^{l} - \left(\mathbf{V}^{T}(\boldsymbol{\theta}^{l})\mathbf{V}(\boldsymbol{\theta}^{l}) + c_l \mathbf{I}_{2K}\right)^{-1}\mathbf{V}^{T}(\boldsymbol{\theta}^{l})\mathbf{b}(\boldsymbol{\theta}^{l}), \quad l = 0, 1, \ldots \tag{54}$$

where

$$\mathbf{V}(\boldsymbol{\theta}) = \mathscr{D}\left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\right)\mathbf{y}^{M} \tag{55}$$

is an $M \times 2K$ matrix (the symbol $\mathscr{D}$ stands for differentiation operation $\mathscr{D} = \frac{\partial}{\partial \boldsymbol{\theta}}$, $\boldsymbol{\theta}^{l}$ denotes the value of the parameter $\boldsymbol{\theta}$ in the *l*-th iteration, $\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}$ is an orthogonal projector onto the orthogonal complement of the matrix $\boldsymbol{\Phi}(\boldsymbol{\theta})$ column space),

$$\mathbf{b}(\boldsymbol{\theta}) = \mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathbf{y}^{M} \tag{56}$$

is an $M \times 1$ vector, $\mathbf{I}_{2K}$ is a $2K \times 2K$ unit matrix, c_l is the Levenberg-Marquardt algorithm constant in the *l*-th iteration.

Levenberg-Marquardt equation (54) is not constructive, it is only a guideline to obtaining iteratively the formant parameter estimates. One can not use this equation directly. In each case of data model, it is necessary to develop (54) computation algorithm in the explicit form using constructive matrix operations (addition, subtraction, multiplication, pseudoinverse, QR decomposition). Data model is described in Section 3.3. One of such unconstructive operations in (54)-(55) is differentiation. In order to implement (55), we need to calculate $\mathscr{D}\left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\right)$ for our data model. This is a rather difficult task. Luckily there exists a formula that simplifies calculation of $\mathscr{D}\left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\right)$. It is shown in (Golub and Pereyra, 1973) that

$$\mathscr{D}\left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\right) = -\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathscr{D}\left(\boldsymbol{\Phi}\right)\mathbf{B} - \left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathscr{D}\left(\boldsymbol{\Phi}\right)\mathbf{B}\right)^{T} \tag{57}$$

where $\mathscr{D}\left(\boldsymbol{\Phi}\right)$ is a three-dimensional tensor formed of $2K$ $M \times 6K$ matrices each containing the partial derivatives of the elements of $\boldsymbol{\Phi}$, and $\mathbf{B}$ is a special generalized inverse of the basis signal matrix $\boldsymbol{\Phi}$.

Denote by $\mathbf{G}_k$ an $M \times 6K$ matrix which is equal to the derivative of $\mathbf{\Phi}$ with respect to the *k*-th component θ_k of the parameter vector $\boldsymbol{\theta}$:

$$\mathbf{G}_k = \frac{\partial \mathbf{\Phi}}{\partial \theta_k}. \tag{58}$$

Then $\mathscr{D}\left(\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})}\right)$ is compounded of 2K $M \times M$ matrices of the form

$$\left(\mathscr{D}\left(\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})}\right)\right)_k = -\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})} G_k \mathbf{B} - (\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})} G_k \mathbf{B})^T. \tag{59}$$

Thus $\mathscr{D}\left(\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})}\right)x$ is formed of $2K$ $M \times 1$ vectors $\left(\mathscr{D}\left(\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})}\right)\right)_k x$ $(i = 1, ..., 6K)$, and hence $V(\theta)$ is an $M \times 2K$ matrix.

The generalized inverse matrix **B** can be calculated by means of a standard *QR*-decomposition of the matrix $\mathbf{\Phi}$. Let **S** stands for a $6K \times 6K$ permutation matrix, $\mathbf{T}_1$ – for a $6K \times 6K$ upper triangular matrix with decreasing diagonal elements, **Q** – for an $M \times M$ orthogonal matrix. Then the matrix **B** is obtained using the formula:

$$\mathbf{B} = \mathbf{S}\left[T_1^{-1} \quad \mathbf{0}_{6K \times (M-6K)}\right]\mathbf{Q}^{\mathrm{T}}. \tag{60}$$

The orthogonal projector $P^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})}$ onto the orthogonal complement of the matrix $\mathbf{\Phi}$ column space can be calculated with a help of the orthogonal matrix **Q** (Golub and Pereyra, 1973):

$$P^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})} = \mathbf{Q}^T \begin{bmatrix} \mathbf{0}_{6K \times 6K} & \mathbf{0}_{6K \times (M-6K)} \\ \mathbf{0}_{(M-6K) \times 6K} & \mathbf{I}_{M-6K} \end{bmatrix} \mathbf{Q}. \tag{61}$$

The matrix $\mathbf{G}_k$ is equal to the partial derivative of the elements of the convoluted matrix $\mathbf{\Phi}$ with respect to the damping factor Λ_k or angular frequency Ω_k of the *k*-th output signal component. Its size is the same as that of

the matrix $\mathbf{\Phi}$, i. e. $M \times 6K$. Denote for $k = 1, ..., K$:

$$\mathbf{G}_{2k-1} = \frac{\partial \mathbf{\Phi}}{\partial \theta_{2k-1}} = \frac{\partial \mathbf{\Phi}}{\partial \Lambda_k}, \qquad \mathbf{G}_{2k} = \frac{\partial \mathbf{\Phi}}{\partial \theta_{2k}} = \frac{\partial \mathbf{\Phi}}{\partial \Omega_k}. \tag{62}$$

It is not difficult to check that:

$$\mathbf{G}_{2k-1} = \frac{\partial \mathbf{\Phi}}{\partial \Lambda_k} = \left[\underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(k-1)}, \mathbf{\Phi}(:, 6k-3), \mathbf{\Phi}(:, 6k-2), \mathbf{\Phi}(:, 6k-1), \mathbf{\Phi}(:, 6k), \right.$$

$$\left. \left\{ \sum_{i=1}^{3} (im-1) e^{(im-1)\Delta t \Lambda_k} \cos(im-1)\Omega_k \right\}_{m=1}^{M}, \left\{ \sum_{i=1}^{3} (im-1) e^{(im-1)\Delta t \Lambda_k} \sin(im-1)\Omega_k \right\}_{m=1}^{M}, \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(K-k)} \right], \tag{63}$$

and

$$\mathbf{G}_{2k} = \frac{\partial \mathbf{\Phi}}{\partial \Omega_k} = \left[\underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(k-1)}, \mathbf{\Phi}(:, 6k-2), \mathbf{\Phi}(:, 6k-3), \mathbf{\Phi}(:, 6k), \mathbf{\Phi}(:, 6k-1), \right.$$

$$\left. -\left\{ \sum_{i=1}^{3} (im-1) e^{(im-1)\Delta t \Lambda_k} \sin(im-1)\Omega_k \right\}_{m=1}^{M}, \left\{ \sum_{i=1}^{3} (im-1) e^{(im-1)\Delta t \Lambda_k} \cos(im-1)\Omega_k \right\}_{m=1}^{M}, \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(K-k)} \right]. \tag{64}$$

Since the columns of the matrix $\mathbf{\Phi}$ and their products with scalars belong to its column space, the products of the projector matrix $\mathbf{P}_{\mathbf{\Phi}(\boldsymbol{\theta})}^{\perp}$ and the

$(6k-5)$-th, $(6k-4)$-th $(6k-3)$-rd, $(6k-2)$-nd columns of the matrices $\mathbf{G}_{2k-1}$ and $\mathbf{G}_{2k}$ are equal to the null vector of length *M*. Thus $\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})}\mathbf{S}_k$ is an $M \times 6K$ matrix whose all columns are zero except the two ones (the $(6k-1)$-st and $(6k)$-th columns).

Parameter estimation algorithm

All the formulas ((56)-(64)) are constructive. Using these formulas, a stepwise algorithm for calculating formant parameter estimates from convoluted data can be present.

Given:

1. The index *k* of the harmonics (in harmonic synthesis method) or formants (in formant synthesis method) under investigation.
2. The sequence **y** obtained from the pitch samples (Selection of the phoneme representative period procedure is presented in the Section 3.6)
3. The initial parameter vector $\boldsymbol{\theta}^0 = \left[\Omega_k^0 \quad \Lambda_k^0\right]^T$.
4. The initial value of Levenberg-Marquardt constant $c_0 = 0.001$.
5. The iteration number $l = 0$.
6. The maximal iteration number $l_{\max}$.
7. The signal estimation relative accuracy in percent $\varepsilon_{\min}$.
8. The initial signal estimation relative accuracy in percent ε_{-1}, for example $\varepsilon_{-1} = 100$.
9. The allowed limit damping factor value $\Lambda_{\lim}$, for example $\Lambda_{\lim} = -0.006$.
10. The allowed maximal Levenberg-Marquardt constant $c_{\max}$, for example $c_{\max} = 10^{10}$.
11. Stop criterion is $\varepsilon_l < \varepsilon_{\min}$ or $l \geq l_{\max}$ or $c_l > c_{\max}$ or $\Lambda_k^l > \Lambda_{\lim}$.

Step 1. Compute the quasipolynomial matrix $\mathbf{\Psi} = \mathbf{\Psi}_k(\boldsymbol{\theta}^l)$ using (31)-(34).

Step 2. Compute the convolution matrix $\mathbf{\Phi} = \mathbf{\Phi}_k(\boldsymbol{\theta}^l)$ by equation (45).

Step 3. Using the standard QR decomposition of the matrix $\mathbf{\Phi}$, find the general inverse matrix **B** according to formula (60).

Step 4. Compute the projector onto the noise subspace $P^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta}^l)}$ by formula (52).

Step 5. Find the projection of **y** onto the noise subspace by formula (56).

Step 6. Determine the error $\varepsilon_l = r_2(\boldsymbol{\theta}^l)/\|y\|^2 \cdot 100\%$, where $r_2(\boldsymbol{\theta}^l)$ is defined by (51).

Step 7. If ($\varepsilon_l < \varepsilon_{l-1}$) then

$$c_l = c_{l-1}/10$$

else

$$c_l = c_{l-1} \cdot 10$$
$$\boldsymbol{\theta}^l = \boldsymbol{\theta}^{l-1}$$

Go to Step 11.

Step 8. Compute the partial derivative of the elements of the matrix $\mathbf{\Phi}$ with respect to Λ_k and Ω_k using formulas (63)-(64).

Step 9. Find the derivative of the projector $\mathscr{D}\left(\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta}^l)}\right)$ according to equation (59).

Step 10. Calculate the matrix $V(\boldsymbol{\theta}^l)$ by formula (55).

Step 11. Compute the parameter vector $\boldsymbol{\theta}^{l+1}$ by (54) and return to Step 1.

END

A comment on this algorithm:

- ✓ The algorithm would be valid in the case when parameters were estimated not for a single formant but for several ones. Practice, however, shows that the results are unstable due to the ill-conditioned matrix $\mathbf{\Psi}$.

3.5. The vowel and semivowel phoneme model in a state space form

In state space representation, the MISO system presented in Fig. 9 can be described as follows (Slivinskas and Šimonytė, 1990):

$$\begin{aligned} x(n+1) &= \boldsymbol{F}x(n) + \boldsymbol{G}u(n) \\ y(n) &= \boldsymbol{H}x(n) \end{aligned} \tag{65}$$

where $\boldsymbol{F}$ is an $6K \times 6K$ block diagonal matrix made of K Jordan blocks $\boldsymbol{F} = \text{blockdiag}(\boldsymbol{F}_1, \dots, \boldsymbol{F}_K)$ with

$$\boldsymbol{F}_k = \begin{bmatrix} a & -b & a & -b & a/2 & -b/2 \\ b & a & b & a & b/2 & a/2 \\ & 0 & a & -b & \text{a} & -\text{b} \\ & & b & a & \text{b} & \text{a} \\ 0 & & & 0 & a & -b \\ & & & & b & a \end{bmatrix} \tag{66}$$

K is the total number of harmonics (in harmonic synthesis method) or formants (in formant synthesis method),

$$\begin{aligned} a &= a_k = e^{\lambda_k \Delta t} cos(2\pi f_k \Delta t) \\ b &= b_k = e^{\lambda_k \Delta t} sin(2\pi f_k \Delta t) \end{aligned} \tag{67}$$

(the subscript k in (66) is omitted for simplicity), $\boldsymbol{G}$ is an $6K \times$ i block diagonal matrix made of K column vectors $\boldsymbol{G} = \text{blockdiag}(\boldsymbol{G}_1, \dots, \boldsymbol{G}_K)$ with

$$\boldsymbol{G}_k = [0, 0, 0, 0, 1, 1]^{\text{T}}, \tag{68}$$

$\boldsymbol{H}$ is an $K \times 8K$ block diagonal matrix made of K row vectors $\boldsymbol{H} = \text{blockdiag}(\boldsymbol{H}_1, \dots, \boldsymbol{H}_{\text{K}})$ with

$$\boldsymbol{H}_{\text{k}} = [A_{k3}\beta_{k3,}\, A_{k3}\gamma_{k3,}\, 0.5A_{k2}\beta_{k2,}\, 0.5A_{k2}\gamma_{k2,}\, 0.5A_{k1}\beta_{k1,} 0.5A_{k1}\gamma_{k1}] \tag{69}$$

where

$$\begin{aligned} \beta_{kl} &= sin(\varphi_{kl}) - cos(\varphi_{kl}) \\ \gamma_{kl} &= sin(\varphi_{kl}) + co\,s(\varphi_{kl}), \end{aligned} \tag{70}$$

$k = 1, \ldots, K, \;\; l = 1, \ldots, 3.$

3.6. Selection of the phoneme representative period

The vowel and semivowel phoneme signals are quasi-periodic, i. e. their periods are not exactly the same. Therefore only a single period for each phoneme is considered. Such a period is usually called a pitch.

Let

$$\boldsymbol{y}^M = [y(0), y(1), y(2), \ldots, y(M-1)]^T \tag{71}$$

be a sequence of samples of the considered pitch of the analysed phoneme. This sequence can be treated as the output of a MISO system (see Section 3.3). In order to automate this selection procedure, the amplitude size as the selection criteria, i. e. the pitch with the highest amplitude is selected as a representative one, is chosen. We are looking for a representative period that is within the 60% of the phoneme signal samples (i.e. 20% of the samples in the signal start part and 20% in the signal end part are rejected). Fig. 10 shows a vowel phoneme signal made of several pitches where the pitch with the highest amplitude is marked with an arrow.

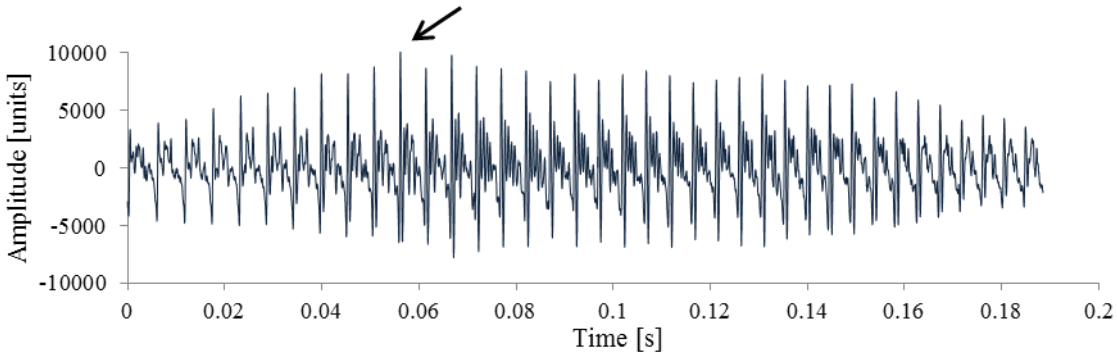


Fig. 10 *Selecting of the vowel phoneme pitch with the highest amplitude*

The representative period is used to compute the parameters of the MISO system impulse response components (25). The start point of the representative pitch (the point t_{start} in Fig. 11) is selected in the following way:

1. The phoneme signal is filtered in the low-pass filter with the $2.5f_0$ bandwidth. Such a bandwidth is chosen in order the filtered signal is the sum of the first two harmonics. This signal approximates the original phoneme signal and its periods coincide with the phoneme signal periods. The filtered signal is shown in Fig. 11.
2. A point of the filtered signal crossing with the abscises axis is chosen (we start from the point whose abscissa coincides with the pitch maximum point abscissa and go to the left along the filtered signal until we find the first point where the filtered signal crosses the x-axis from the bottom).
3. The pitch start point is searched in the neighbourhood of the found crossing point. In the beginning a point is looked for with a negative amplitude of the smallest absolute value in the vicinity of the first $f_s/3f_0$ points to the right from the crossing point or the point where the phoneme signal crosses the abscises axis. If such a point is found, the first point to the left with a negative amplitude of the smallest absolute value in the vicinity of the first $f_s/3f_0$ points or the point of the signal crossing with the abscises axis is looked for. If such a point does not exist, then a point of the filtered signal crossing with the abscises axis is considered as the pitch beginning point.

The procedure of selection of the end point of the representative pitch (the point t_{end} in Fig. 11) is as follows. The pitch start point is selected and the point that is at the distance of $T_0 = 1/f_0$ from this start point is found. With a help of the found point, the pitch end point is determined in the same way as is described above (with the start point).

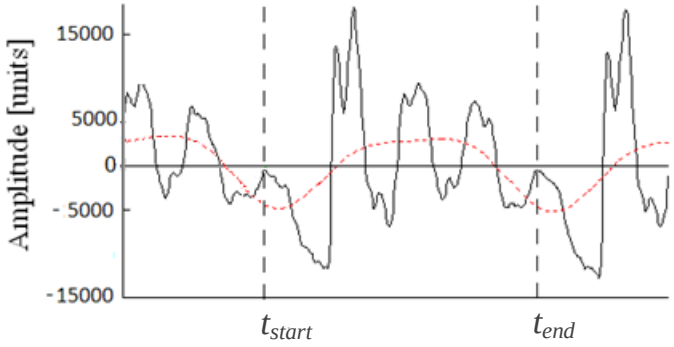


Fig. 11 *Selecting of the start and end points of the representative pitch*

3.7. Determining of the inputs

A MISO system proposed in Section 3.3 for vowel phoneme modelling is stationary. We compute parameters of the impulse responses of the SISO components of this system under the assumption that the unit impulse is given to the system input. If the unit impulses are given to the system input at intervals $T = 1/f_0$ (see Fig. 12), then a signal with equal periods is obtained in the system output; those periods are the same as the representative period.

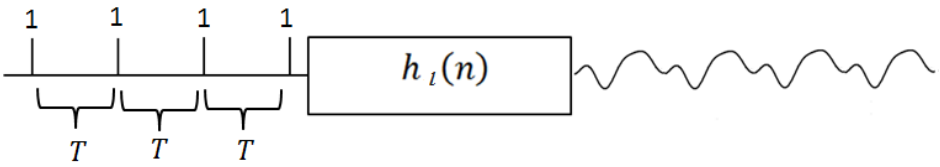


Fig. 12 *A SISO system with the unit impulse inputs*

Such a signal sounds synthetically. In order to obtain a quasiperiodic output signal, the system should be excited by impulses with different amplitudes. A procedure of determining such impulses is presented in this section.

First it is necessary to divide (segment) automatically the whole phoneme signal into periods. The segmentation procedure determines the start and end points of each period in the same way as in the case of the representative pitch (see Section 3.6). An example of such segmentation is shown in Fig. 13.

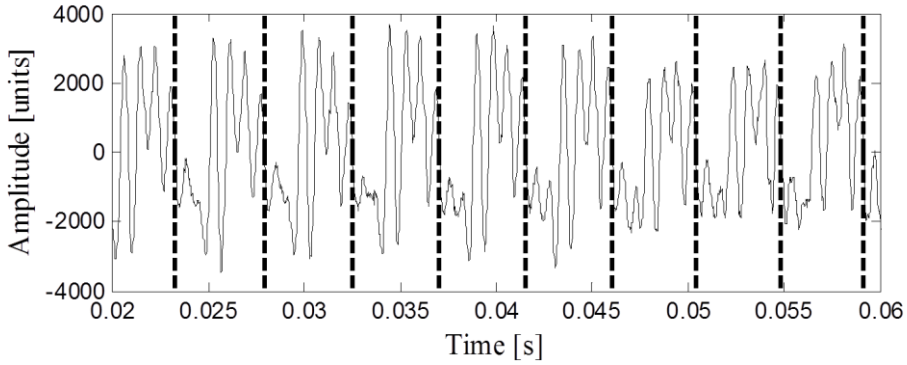


Fig. 13 *A part of the phoneme /a/ signal divided into periods*

After segmenting a signal into periods, the start point of each period is determined. Then the first harmonic or formant (harmonic – in harmonic synthesis method, formant – in formant synthesis method) component is divided into time intervals using the determined points. In each of these intervals, the maximum points are found. The ordinates of these points are stored in a vector that is the first column of the matrix that is called the input matrix. Analogously, the second component is segmented into the same time intervals. Then again the maximum point in each of these intervals is founded. The ordinates of these points are stored in a vector that becomes the second column of the input matrix. The algorithm is continued until all the L input matrix columns are filled in. In the end, we get a $P \times L$ matrix where P is the number of time intervals .

The amplitude of the impulse given to the system input whose output signal is the representative period must be equal to one. Therefore we have to norm the input matrix. The norming is carried out separately for each column. The norming procedure is as follows: a row that corresponds to the representative pitch is selected and all the values of that column by the value at intersection of this column and the selected row are divided. After the norming procedure is completed, the ratios of the representative period amplitude and amplitudes of the all periods are obtained. These ratios determine the dynamics of the real harmonic (formant) component amplitudes. The inputs of the first

three channels of a MISO system are presented in Fig. 14 (the harmonic synthesis method case).

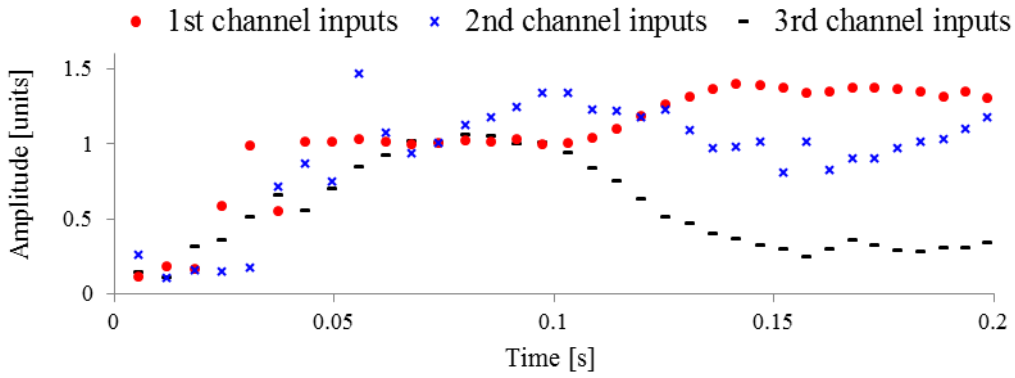


Fig. 14 *The inputs of the first three channels of a MISO system*

From Fig. 14 we see that the system inputs are changing in time. For example, in the second half of the inputs the 1st channel impulse amplitudes increase, while the 3rd channel impulse amplitudes reduce significantly.

The input amplitudes are described by parabolas. Each parabola is characterized by the following three parameters: $\{y_{max}, t_{max}, t_{lim}\}$, where y_{max} is the maximum input value, t_{max} – the time instant of the maximum input value, t_{lim} – the phoneme length.

The parabola describes the input amplitudes in time instants $t = 0, T_1, T_2, \dots, t_{lim}$.

To these inputs two additional input sequences are added. The argument of the first sequence is $[t_{start}, 0]$, and the argument of another is $[t_{lim}, t_{end}]$. These sequences are also described by parabolas. The first parabola depicts the input growth from zero to the initial value, while the latter – the input damping until the zero value. Fig. 15 shows the total input curve.

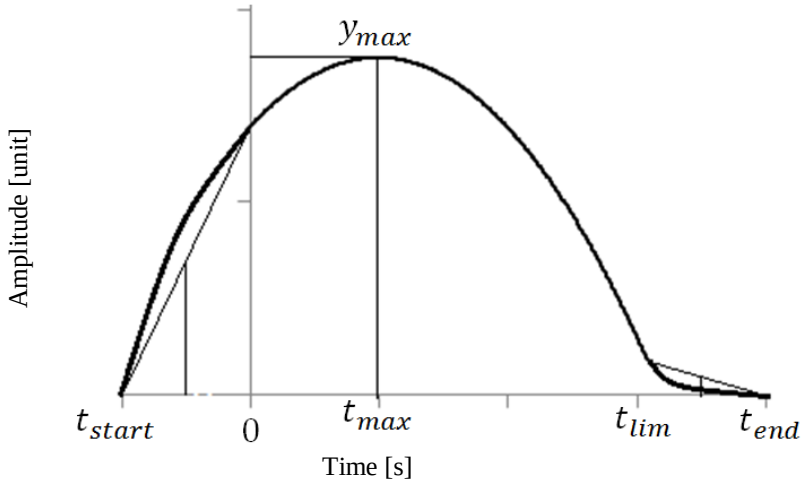


Fig. 15 *The total input curve*

Since we fix the start point of each period, it is not difficult to calculate the lengths of the phoneme periods:

$$T = [T_1, T_2, \dots, T_P]. \tag{72}$$

The entries of the vector T define the distances between the input impulses.

3.8. Conclusions of Section 3

The assumption that the impulse response of the SISO system corresponding to a single formant decays after three fundamental periods is used. This assumption allowed us to apply Levenberg-Marquardt method to estimate parameters from convoluted data.

A new fundamental frequency refining algorithm is proposed.

A new method that allows one to select the representative period automatically is given.

In the case when the unit impulses are inputted to the system, the output signal is periodic with identical periods. This period identity is the main reason of unnatural (synthetic) sounding of the output signal. In order for the synthesized signal to sound more naturally, impulses of different amplitudes

and periods as inputs instead of the unit impulses with a constant period are used.

The input sequence of each phoneme has been described by three parabolas. The first parabola characterized the slowdown growth of the formant, the second parabola described the main time region, and the third one characterized the slowdown decreasing of the formant. The inputs for each phoneme are arranged consequently with overlapping.

4

Experimental research

Real data in this section are consider. This data are samples of natural sounds. The sounds were recorded using a microphone and the "Sound Record" program to a .wav file of the audio format with the following parameters: PCM 48 kHz; 16 bit; stereo. This frequency corresponds to the sampling interval of 21 μs. The experiments were carried out using my programs developed in MATLAB and C#.

4.1. Fundamental frequency estimation using the MUSIC method

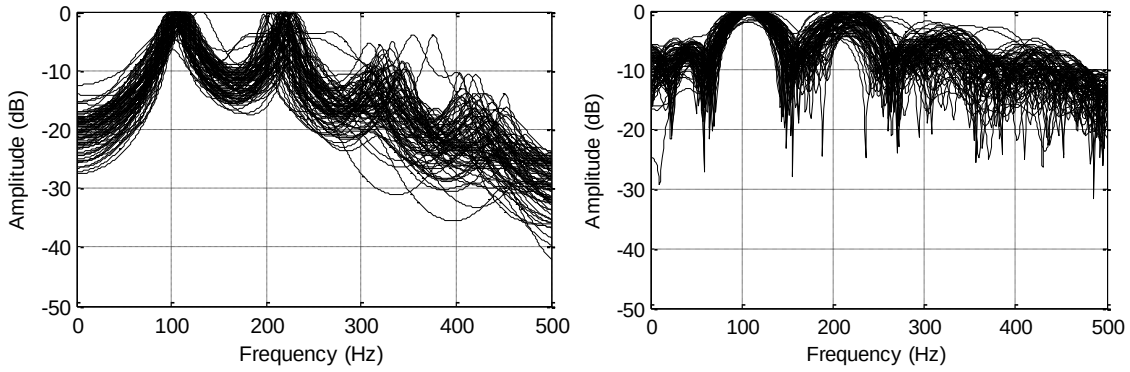
The MUSIC method and DFT (Discrete Fourier Transform) method for Lithuanian male vowels /a/, /i/, /o/, /u / were applied. The MUSIC spectrum was calculated using the formula:

$$\mathrm{MUSIC}(f) = 10 \cdot \lg\left(\left|\hat{P}_{MU}(e^{j2\pi f})\right|\right) \tag{73}$$

where $\hat{P}_{MU}(e^{j2\pi f})$ is defined by (9).

For each of the vowels mentioned above, 80 records of length 1024 points

were considered. For each of these records, the spectra and estimates of the fundamental frequency by the MUSIC method and DFT are calculated, and their mean $E(\hat{f}_0)$ and standard deviation $\sigma(\hat{f}_0)$ were obtained. The results are shown in Figure 16 and Table 7.



a) MUSIC spectrum. b) FFT spectrum.

Fig. 16 *Spectrum estimates for a Lithuanian male vowel /u/ for 80 speech signal realisations*

Table 7 *The mean and standard deviation of the fundamental frequency estimates obtained by the MUSIC method and DFT method*

Vowel / **Characteristics**	**/a/**	**/i/**	**/o/**	**/u/**
The mean $E(\hat{f}_0)_{MUSIC}$	112.00	114.75	108.20	107.35
The mean $E(\hat{f}_0)_{DFT}$	124.80	127.15	118.95	120.12
The standard deviation $\sigma(\hat{f}_0)_{MUSIC}$	3.55	2.36	4.13	4.62
The standard deviation $\sigma(\hat{f}_0)_{DFT}$	5.59	4.18	4.18	5.07

We see from Table 7 that the difference between the estimated frequencies of the vowels /o/ and /u/, however, is small – about 1 Hz (0.85 Hz for the MUSIC, and 1.17 Hz for the DFT). The smallest standard deviation 2.36 Hz was obtained by the MUSIC method for the vowel /i/, and the largest – 5.59 Hz – by the DFT method for the vowel /a/. It is easy to notice that the DFT standard deviation values are higher than those of the MUSIC method for

all vowels.

Since the DFT and MUSIC methods give different estimates of the fundamental frequency, which estimate describes the real situation more accurately have to be checked. For each vowel, a model of the sum of ten harmonics where the first harmonic frequency was the fundamental frequency estimate (the DFT or MUSIC) were used. The parameters of the harmonics were estimated by a usual linear least squares method. The relative estimation errors are shown in Table 8.

Table 8 *The relative approximation error of the vowel signals by the sum of ten harmonics using the fundamental frequency estimates obtained by the MUSIC method and DFT method*

Vowel / **Error**	**/a/**	**/i/**	**/o/**	**/u/**
err_{MUSIC}	37.32%	44.74 %	22.99%	8.86 %
err_{DFT}	78.75%	54.70%	68.33 %	37.94 %

We see from Table 8 that the errors obtained with the MUSIC fundamental frequency estimate are smaller than those obtained with the DFT fundamental frequency estimate. The fact that the errors are rather large can be attributed to complexity of the real sound signals. Harmonics of such signals are time variant, and it is very difficult to describe the signal with time invariant frequency models.

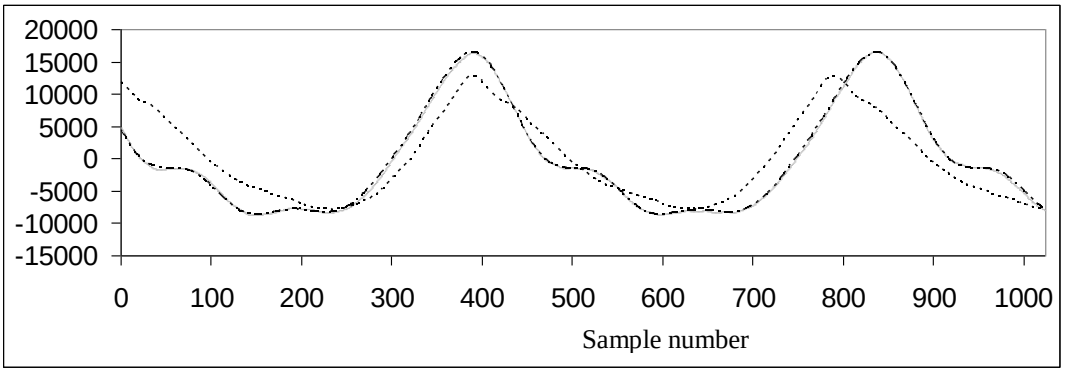


Fig. 17 *The true and estimated speech signal of the vowel /u/ (solid line – the true speech signal, dotted line – the estimated signal (DFT method), dash-dotted line – the estimated signal (MUSIC method))*

The true signal of the vowel /u/ and its estimates obtained by the DFT and MUSIC methods are shown in Figure 17. The signal estimates are obtained using a model of the sum of 10 harmonics. One can see that the MUSIC estimate almost coincides with the true signal.

4.2. Vowels and semivowels modelling by formant and harmonic methods

The modelling for all the vowel and semivowel phonemes using 50 utterances by female and 50 utterances by male are carried out. The synthesized sounds obtained by the harmonic synthesis method are compared with those obtained by the formant method. Lists of Lithuanian words used in the experiment are presented in Table 1 of Section 2.5.1 and in Table 2 of Section 2.5.2. In order to estimate the model quality, the average spectrum are calculated. The comparison of the spectra of the true phoneme /a/ signal and its models is shown in Fig. 18.

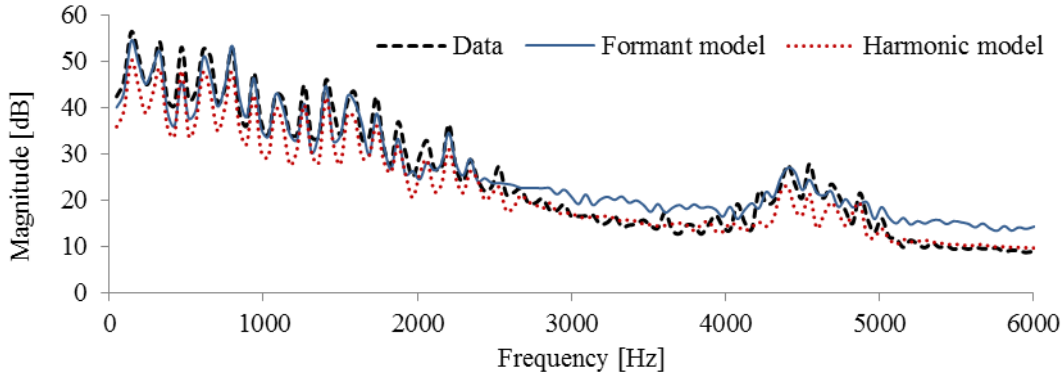


Fig. 18 *The spectra of the true phoneme /a/ signal and its models*

Fig. 18 shows that the obtained spectra almost coincide. The audio test revealed that the differences have no significant influence to the sound intelligibility and quality.

The average RMSE of the estimated signal spectrum and its confidence intervals for each of the 28 vowel and 19 semivowel phonemes are presented in Table 9 and Table 10. The confidence intervals are stated at the 95 % confidence level.

Table 9 *The average RMSE and its confidence intervals for the estimated vowel phoneme signal spectrum*

Phoneme	Formant method case				Harmonic method case			
	Female phoneme		Male phoneme		Female phoneme		Male phoneme	
	RMSE	Confidence intervals	RMSE	Confidence intervals	RMSE	Confidence intervals	RMSE	Confidence intervals
/a/	13.0 %	[10.8, 14.8]	13.1 %	[10.7, 14.9]	12.7 %	[11.3, 14.6]	12.4 %	[11.1, 14.5]
/a`/	14.0 %	[11.8, 15.8]	12.7 %	[11.6, 14.2]	12.4 %	[10.8, 13.5]	11.9 %	[9.2, 13.1]
/a:/	12.9 %	[11.7, 14.1]	13.5 %	[11.6, 15.6]	12.4 %	[10.8, 12.9]	12.8 %	[10.9, 14.1]
/a:´/	12.9 %	[10.8, 15.9]	11.6 %	[10.2, 12.8]	10.4 %	[9.4, 11.1]	11.3 %	[10.1, 12.9]
/a:˜/	13.7 %	[12.0, 15.5]	14.9 %	[11.4, 17.5]	13.5 %	[11.6, 15.4]	13.8 %	[11.9, 15.1]
/e/	14.1 %	[12.4, 15.7]	13.7 %	[12.0, 15.5]	13.6 %	[11.2, 14.9]	13.4 %	[11.4, 15.1]
/e`/	13.9 %	[12.1, 15.8]	13.9 %	[12.4, 15.2]	13.0 %	[11.0, 14.8]	12.9 %	[11.1, 14.8]
/e:/	11.6 %	[10.6, 12.7]	14.1 %	[12.9, 15.8]	9.7 %	[8.3, 10.8]	11.3 %	[10.0, 12.5]
/e:´/	11.3 %	[10.0, 12.6]	14.8 %	[14.0, 15.7]	8.5 %	[7.3, 9.1]	10.9 %	[9.1, 12.8]
/e:˜/	14.2 %	[12.6, 15.8]	13.8 %	[12.4, 15.9]	12.1 %	[11.5, 13.7]	13.1 %	[11.4, 14.8]
/ė:/	12.8 %	[11.0, 14.3]	14.4 %	[13.8, 15.5]	12.4 %	[10.4, 14.4]	11.8 %	[10.0, 13.2]
/ė:´/	13.6 %	[11.9, 15.5]	12.9 %	[11.7, 14.2]	12.8 %	[10.9, 14.0]	12.2 %	[10.5, 13.7]
/ė:˜/	16.8 %	[16.0, 19.0]	13.5 %	[12.3, 15.8]	12.5 %	[10.9, 14.1]	13.0 %	[11.4, 14.2]

/i/	13.1 %	[10.7, 14.9]	14.6 %	[13.8, 15.4]	13.3 %	[11.5, 14.8]	12.9 %	[10.8, 15.9]
/i`/	12.8 %	[11.1, 14.4]	13.2 %	[10.8, 14.9]	12.4 %	[10.8, 13.9]	12.1 %	[9.9, 14.8]
/i:/	13.6 %	[11.9, 15.5]	12.8 %	[10.7, 15.8]	12.3 %	[10.8, 13.9]	12.9 %	[11.3, 14.8]
/i:´/	12.9 %	[10.7, 15.9]	15.9 %	[14.4, 16.5]	13.0 %	[11.5, 14.2]	12.7 %	[9.9, 14.3]
/i:˜/	13.9 %	[12.5, 15.2]	15.6 %	[11.8, 19.3]	12.3 %	[10.9, 13.8]	12.4 %	[10.7, 14.8]
/o/	11.8 %	[10.8, 12.8]	14.9 %	[14.1, 15.9]	11.1 %	[9.7, 13.5]	10.9 %	[8.9, 13.8]
/o`/	11.7 %	[10.2, 12.9]	12.8 %	[10.9, 14.4]	9.2 %	[8.0, 9.6]	9.9 %	[8.6, 11.7]
/o:/	14.1 %	[11.8, 15.9]	13.1 %	[12.7, 18.8]	12.4 %	[10.4, 16.9]	12.1 %	[9.8, 13.8]
/o:´/	12.7 %	[10.9, 14.2]	13.5 %	[12.2, 15.7]	12.3 %	[10.5, 15.9]	12.4 %	[11.7, 13.7]
/o:˜/	14.0 %	[12.2, 15.3]	16.6 %	[12.8, 20.3]	11.6 %	[11.0, 12.4]	13.6 %	[11.8, 15.0]
/u/	13.4 %	[11.6, 15.4]	14.5 %	[13.7, 14.7]	13.2 %	[11.3, 15.0]	14.3 %	[13.8, 15.2]
/u`/	13.5 %	[11.8, 15.3]	12.9 %	[12.5, 14.2]	12.5 %	[10.9, 16.8]	10.3 %	[9.0, 11.6]
/u:/	14.6 %	[11.2, 17.1]	14.7 %	[14.0, 15.9]	13.3 %	[11.9, 15.1]	14.4%	[12.8, 16.2]
/u:´/	13.5 %	[11.8, 15.5]	18.1 %	[17.0, 19.8]	13.2 %	[12.1, 14.9]	13.7 %	[11.8, 15.1]
/u:˜/	16.4 %	[14.4, 18.8]	19.6 %	[18.3, 20.7]	15.4 %	[12.8, 16.7]	15.8 %	[14.0, 18.3]
Average	13.5 %		14.3 %		12.3 %		12.5 %	

Table 10 *The average RMSE and its confidence intervals for the estimated semivowel phoneme signal spectrum*

Phoneme	**Formant method case**				**Harmonic method case**			
	Female phoneme		**Male phoneme**		**Female phoneme**		**Male phoneme**	
	RMSE	**Confidence intervals**	**RMSE**	**Confidence intervals**	**RMSE**	**Confidence intervals**	**RMSE**	**Confidence intervals**
/j"/	16.2 %	[13.5, 17.4]	21.8 %	[19.1, 23.2]	15.8 %	[15.3 16.8]	15.9 %	[14.8, 17.9]
/l/	17.2 %	[14.5, 18.4]	26.7 %	[25.6, 27.6]	17.6 %	[16.5, 19.6]	18.9 %	[16.2, 21.2]
/l˜/	18.5 %	[17.6, 19.7]	22.6 %	[20.4, 24.2]	22.4 %	[21.5, 24.0]	20.5 %	[19.5, 21.9]
/l"/	25.9 %	[25.5, 28.3]	27.5 %	[25.7, 31.0]	24.3 %	[23.1, 26.6]	25.4 %	[23.6, 26.9]
/l"˜/	25.7 %	[24.6, 27.2]	17.7 %	[17.0, 18.2]	23.4 %	[22.4, 24.6]	19.1 %	[16.9, 20.7]
/m/	16.5 %	[15.5, 18.0]	19.1 %	[17.1, 20.8]	15.0 %	[14.7, 15.5]	15.4 %	[14.3, 17.9]
/m˜/	17.7 %	[16.0, 19.2]	18.7 %	[17.6, 19.2]	11.8 %	[10.3, 13.0]	12.2 %	[10.5, 13.5]
/m"/	10.9 %	[7.9, 13.3]	17.9 %	[16.8, 18.9]	9.1 %	[7.2, 10.8]	12.1 %	[9.9, 13.8]
/m"˜/	19.7 %	[17.8, 20.9]	14.4 %	[11.0, 16.9]	15.7 %	[14.6, 18.5]	12.9 %	[10.9, 15.3]
/n/	14.6 %	[11.1, 17.1]	11.9 %	[8.9, 13.9]	11.8 %	[9.2, 14.0]	10.9 %	[7.9, 13.4]
/n˜/	19.1 %	[16.0, 20.9]	17.6 %	[16.9, 17.8]	15.5 %	[12.2, 18.1]	14.8 %	[11.3, 17.3]
/n"/	23.2 %	[20.6, 25.2]	16.7 %	[14.7, 18.9]	16.3 %	[15.0, 17.8]	14.7 %	[13.6, 15.7]
/n"˜/	25.9 %	[25.2, 26.7]	22.6 %	[20.8, 23.8]	21.7 %	[21.1, 22.3]	18.9 %	[17.7, 19.3]
/r/	18.6 %	[18.1, 18.9]	23.8 %	[21.4, 26.4]	16.7 %	[14.9, 17.9]	17,6 %	[16.8, 18.2]
/r˜/	22.8 %	[20.1, 24.2]	25.1 %	[24.7, 26.9]	18.0 %	[16.9, 20.3]	19.1 %	[16.9, 20.9]
/r"/	18.7 %	[18.2, 18.9]	23.3 %	[21.7, 25.1]	15.1 %	[12.5, 18.6]	17.2 %	[15.1, 18.6]
/r"˜/	23.0 %	[21.6, 24.9]	24.1 %	[23.6, 25.8]	21.1 %	[18.6, 24.1]	20.9 %	[19.5, 22.3]
/v/	19.5 %	[16.4, 21.2]	16.6 %	[14.5, 18.9]	15.1 %	[13.8, 16.4]	14.7 %	[11.9, 17.0]
/v"/	16.4 %	[14.3, 18.9]	19.6 %	[16.4, 21.4]	15.4 %	[13.7, 16.9]	13.6 %	[12.1, 15.6]
Average	19.5 %		20.4 %		16.9 %		16.5 %	

The graphical representation of the average RMSE shown in Table 9 and Table 10 is presented in Fig. 19 and in Fig. 20.

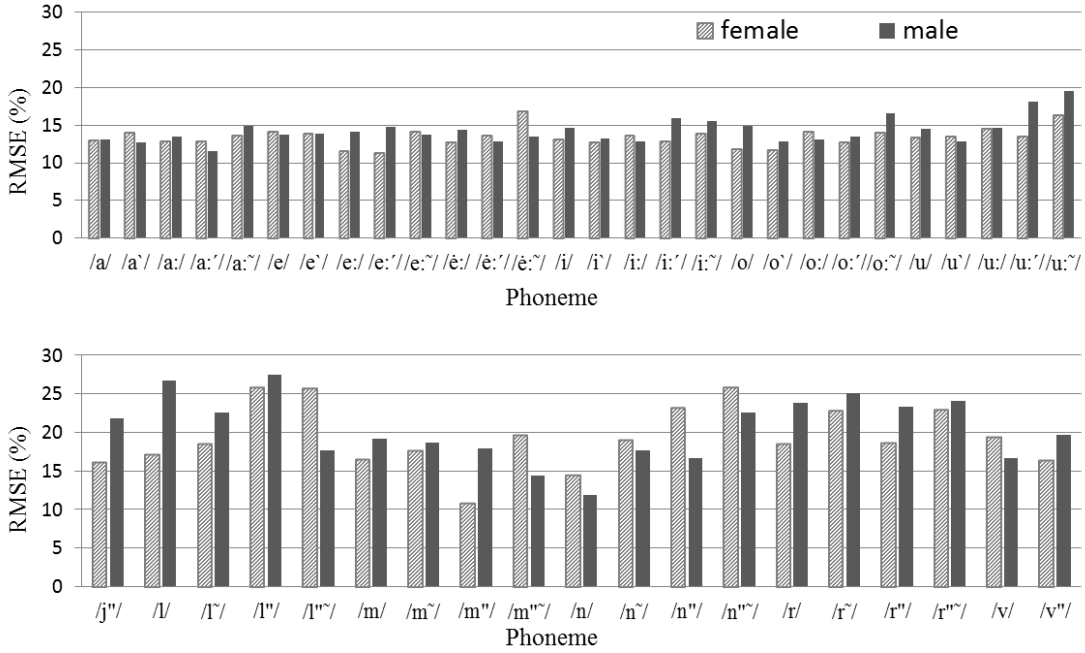


Fig. 19 *The average RMSE for the estimated signal spectrum (formant method case): the upper plot – vowel phonemes, the lower plot - semivowel phonemes*

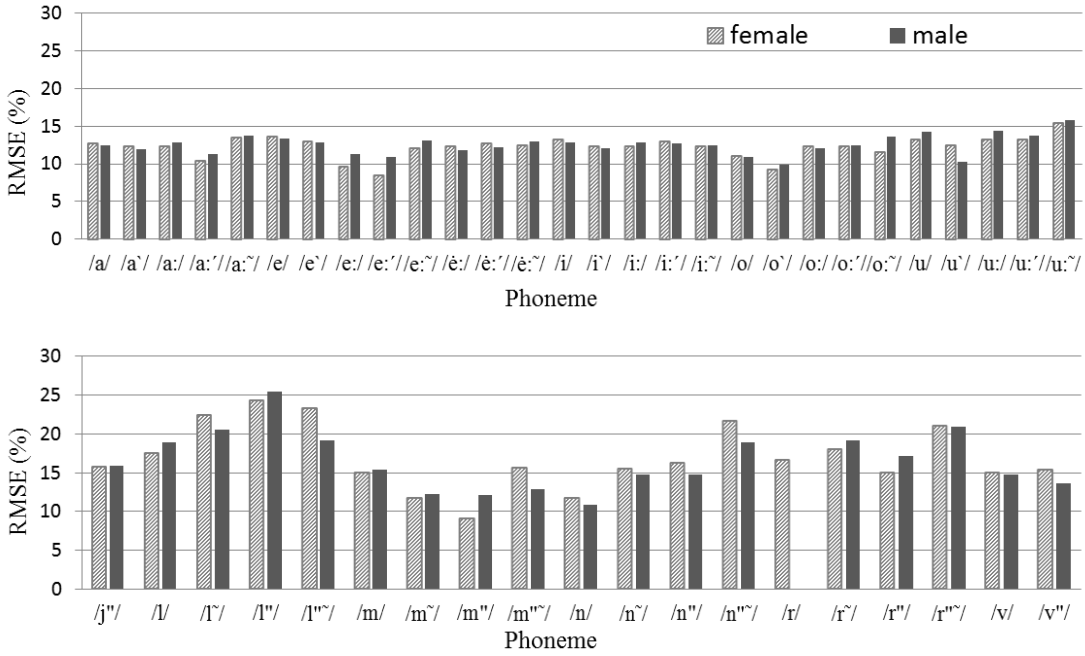


Fig. 20 *The average RMSE for the estimated signal spectrum (harmonic method case): the upper plot – vowel phonemes, the lower plot - semivowel phonemes*

The average RMSE for the estimated signal spectrum for all male and female vowels is equal to 13.9 % in the formant method case and 12.4 % in the harmonic method case. The average RMSE for the estimated signal spectrum for all male and female semivowels is equal to 19.9 % in the formant method case and 16.7 % in the harmonic method case.

Both the spectrum estimation errors and audio test revealed that the quality difference between the sounds synthesized by the harmonic and formant methods is small.

The computation complexity of the proposed algorithms has been evaluated. The average time of the phoneme parameters estimation and average time of the phoneme synthesis for each of the 28 vowel and 19 semivowel phonemes are presented in Table 11 and Table 12.

Table 11 *The average time of the vowel phoneme parameters estimation and the vowel phoneme synthesis (time measured in second)*

Phoneme	Formant method case				Harmonic method case			
	Female phoneme		Male phoneme		Female phoneme		Male phoneme	
	Time of parameter estimation	Time of synthesis	Time of parameter estimation	Time of synthesis	Time of parameter estimation	Time of synthesis	Time of parameter estimation	Time of synthesis
/a/	15	0.07	19	0.09	49	0.3	52	0.5
/a`/	19	0.14	16	0.14	88	0.6	72	0.9
/a:/	22	0.07	19	0.8	87	0.4	88	0.7
/a:´/	32	0.08	33	0.09	72	0.3	69	0.5
/a:˜/	19	0.16	15	0.17	64	0.5	65	0.9
/e/	15	0.05	16	0.09	23	0.2	30	0.4
/e`/	23	0.06	27	0.05	46	0.3	49	0.5
/e:/	22	0.09	22	0.08	30	0.3	36	0.4
/e:´/	25	0.12	23	0.13	45	0.6	45	0.8
/e:˜/	30	0.08	31	0.11	46	0.4	47	0.6
/ė:/	15	0.08	16	0.07	30	0.3	32	0.5
/ė:´/	21	0.12	18	0.11	39	0.5	39	0.8
/ė:˜/	19	0.11	20	0.12	51	0.5	51	0.9
/i/	18	0.06	17	0.05	28	0.2	33	0.3
/i`/	14	0.08	14	0.08	30	0.2	34	0.4
/i:/	7	0.10	4	0.12	22	0.3	20	0.5

/i:´/	10	0.14	11	0.12	38	0.5	41	0.9
/i:˜/	24	0.09	23	0.08	42	0.3	45	0.5
/o/	19	0.12	21	0.11	30	0.4	32	0.6
/o`/	20	0.13	18	0.12	56	0.5	54	0.8
/o:/	16	0.07	18	0.07	28	0.3	28	0.5
/o:´/	18	0.10	15	0.11	34	0.4	33	0.7
/o:˜/	12	0.11	13	0.12	44	0.4	48	0.7
/u/	15	0.04	17	0.08	32	0.2	35	0.4
/u`/	25	0.06	27	0.07	31	0.3	30	0.5
/u:/	11	0.07	9	0.06	27	0.2	26	0.3
/u:´/	15	0.08	18	0.07	31	0.4	31	0.6
/u:˜/	18	0.08	15	0.08	37	0.4	36	0.6

Table 12 *The average time of the semivowel phoneme parameters estimation and the semivowel phoneme synthesis (time measured in second)*

Phoneme	**Formant method case**				**Harmonic method case**			
	Female phoneme		**Male phoneme**		**Female phoneme**		**Male phoneme**	
	Time of parameter estimation	**Time of synthesis**	**Time of parameter estimation**	**Time of synthesis**	**Time of parameter estimation**	**Time of synthesis**	**Time of parameter estimation**	**Time of synthesis**
/j"/	9	0.10	8	0.11	20	0.3	21	0.5
/l/	11	0.04	12	0.04	33	0.1	31	0.2
/l˜/	8	0.05	9	0.06	25	0.1	24	0.2
/l"/	12	0.02	12	0.04	37	0.1	24	0.2
/l"˜/	13	0.03	15	0.04	29	0.2	31	0.3
/m/	8	0.04	8	0.04	26	0.2	29	0.4
/m˜/	11	0.11	8	0.12	29	0.6	30	0.8
/m"/	8	0.06	12	0.06	21	0.3	19	0.5
/m"˜/	16	0.12	14	0.11	39	0.6	43	0.8
/n/	13	0.08	14	0.07	29	0.4	29	0.6
/n˜/	17	0.14	22	0.13	47	0.7	30	1.1
/n"/	8	0.07	14	0.06	22	0.4	36	0.7
/n"˜/	10	0.03	14	0.05	33	0.2	22	0.3
/r/	16	0.08	12	0.09	34	0.4	26	0.5
/r˜/	14	0.06	14	0.06	29	0.3	30	0.5
/r"/	15	0.04	14	0.04	29	0.1	31	0.1
/r"˜/	18	0.05	22	0.06	31	0.2	30	0.2
/v/	11	0.08	14	0.08	23	0.4	37	0.6
/v"/	12	0.06	14	0.05	24	0.3	36	0.5

The average time of the phoneme parameter estimation for all the male and female vowels and semivowels is equal to 16.1 s in the formant method case

and 37.2 s in the harmonic method case. The average time of the phoneme synthesis for all the male and female vowels and semivowels is equal to 0.09 s in the formant method case and 0.44 s in the harmonic method case.

4.3. Diphthong modelling

A part of utterance of the Lithuanian word "laimė" („happiness") corresponding to the compound diphthong /ai/ is considered. The duration of this part was 0.47 s. Let $\mathbf{y} = (y_1, y_2, ..., y_N)$ be equidistant samples of the diphthong /ai/ where N = $0.47 \cdot 48\ 000$ = **22 560** samples. These samples are shown in Fig. 21. It is not difficult to see that this discrete signal exhibits a relative periodicity. One can count 44 periods in total. The first periods belong to the vowel /a/, the last – to the vowel /i/. The middle periods represent transition from the first vowel to the second one.

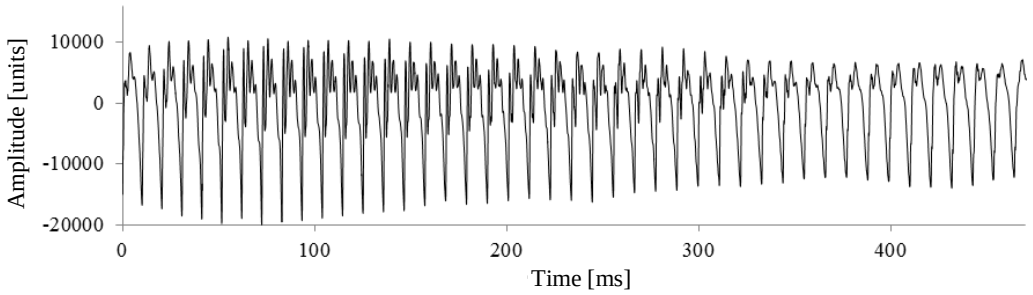
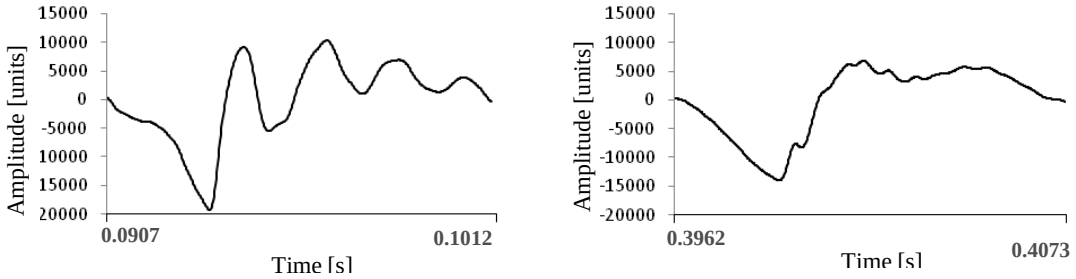


Fig. 21 *The samples of a discretized version of the diphthong "ai" of the Lithuanian word "laimė"*

Two periods corresponding to the vowel /a/ and the vowel /i/ are select. The selected periods are called pitches. The pitch corresponding to the vowel /a/ is the 10-th period of the diphthong /ai/, and the pitch corresponding to the vowel /i/ is the 38-th period. These pitches are shown in Fig. 22.



a) b)

Fig. 22 *The pitch corresponding a) to the vowel /a/ and b) to the vowel /i/*

One can see that the pitch corresponding to the vowel /a/ has four "teeth" while the pitch of the vowel /i/ has two shorter "teeth".

The magnitude responses of the vowels /a/ and /i/ are presented in Fig. 23.

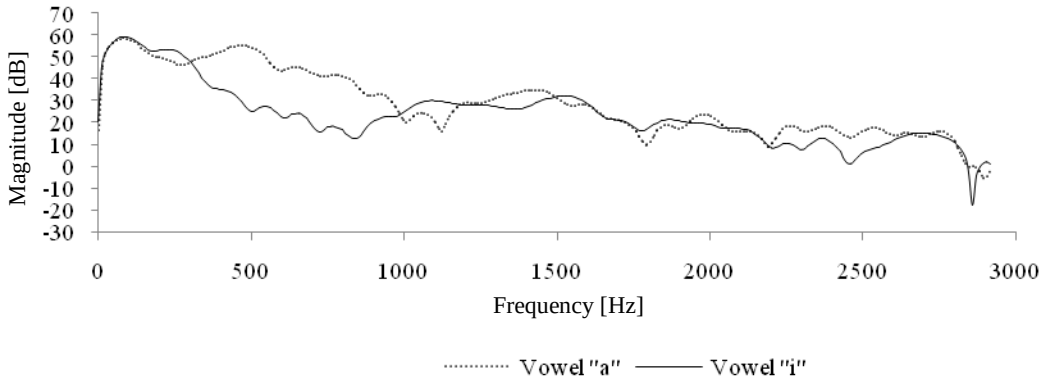


Fig. 23 *The magnitude response of the vowels "a" and "i"*

In the frequency range of 700-950 Hz, the magnitude of the vowel /i/ decreases almost ten times, i. e. by 20 dB. That is a distinctive feature of this vowel. Another distinctive feature is a peak in the range of 950-1100 Hz.

After analysing the magnitude responses of the vowels /a/ and /i/, we selected 19 formant intervals (regions) for the vowel /a/, and 21 intervals for the vowel /i/. The procedure of selection of these intervals is as follows: first, we choose a peak of the magnitude response and go down along this response to the left from the peak until we reach the nearest local minimum. The frequency corresponding to this minimum is the start point of the formant

interval. The end point is obtained analogously going down to the right from the peak. The formant intervals for the vowels /a/ and /i/ are shown in Table 13.

In each of these intervals we carried out the inverse Fourier transform. We obtained 21 signals of length 504 points for the vowel /a/, and 19 signals of length 532 points for the vowel /i/. For each of these signals, we estimated parameters of the quasipolynomial model (25). Estimation was done using parameter estimation algorithm described in Section 3.4.

Table 13 *Formant intervals for the vowels /a/ and /i/*

Interval number	Formant intervals for the vowel /a/	Formant intervals for the vowel /i/
1	30 – 200 Hz	30 – 180 Hz
2	201 – 265 Hz	181 – 440 Hz
3	266 – 380 Hz	441 – 545 Hz
4	381– 600 Hz	546 – 645 Hz
5	601 – 740 Hz	646 – 760 Hz
6	741– 885 Hz	761– 870 Hz
7	886 – 1000 Hz	871 – 990 Hz
8	1001–1105 Hz	991 – 1170 Hz
9	1106 – 1220 Hz	1171 – 1300 Hz
10	1221– 1430 Hz	1301– 1475 Hz
11	1431 – 1550 Hz	1476 – 1650 Hz
12	1551 – 1650 Hz	1651 – 1770 Hz
13	1651 – 1785 Hz	1771 – 1885 Hz
14	1786 – 1890 Hz	1886 – 2020 Hz
15	1891 – 2075 Hz	2021 – 2190 Hz
16	2076 – 2180 Hz	2191 – 2285 Hz
17	2181 – 2300 Hz	2286 – 2410 Hz
18	2301 – 2445 Hz	2411 – 2590 Hz
19	2446 – 2600 Hz	2591 – 2810 Hz
20	2601 – 2675 Hz	
21	2676 – 2820 Hz	

The procedure of the quasipolynomial model obtaining from the data of a single formant interval will be described in more detail. Consider the 3-rd

formant interval of the vowel /a/. First, the initial estimates of the damping factor and angular frequency: $\Lambda_0 = -0.02$, $\Omega_0 = 0.046$ [rad/sample] are selected. These estimates in the iterative procedure described by (54) where $\boldsymbol{\theta} = [\Lambda, \Omega]^T$ are used. This procedure is repeated until the number of iterations is less than 100 or the estimation error is less than 0.5. The following estimates $\Lambda = -0.0095$, $\Omega = 0.044$ [rad/sample] are obtained.

The model and the data for the 3-rd formant interval are shown in Fig. 24.

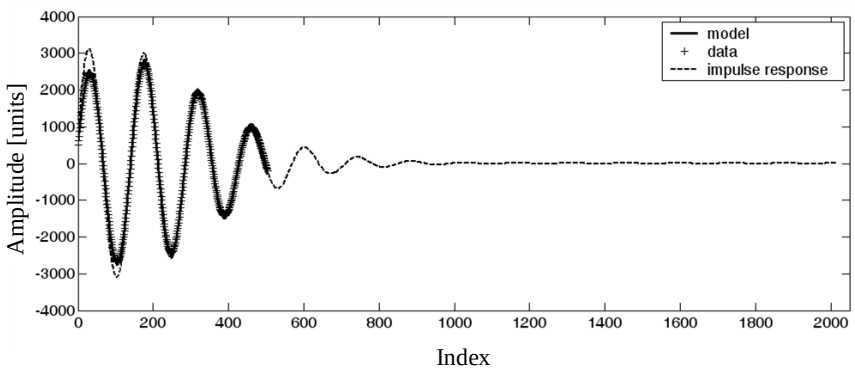


Fig. 24 *The data and estimated model for the 3-rd formant interval*

The root-mean-square estimation error is equal to 3.93 %.

Quasipolynomial parameters of the vowels /a/ and /i/ for all the formant intervals are presented in Table 14 and Table 15, respectively.

Table 14 *Formant parameters of the vowels /a/*

Formant number k	Frequency [Hz] f_k	Damping [unit] λ_k	Amplitude1 [unit] A_{1k}	Amplitude2 [unit/s] A_{2k}	Amplitude3 [unit/s2] A_{3k}	Phase1 [rad] φ_{1k}	Phase2 [rad] φ_{2k}	Phase3 [rad] φ_{3k}
1	73	-932	2481720	24388	77.88	-0.001	-1.784	2.572
2	235	-334	2184.06	9.29	0.09	1.229	2.001	1.027
3	339	-458	3460.81	17.09	0.36	0.213	0.946	-0.189
4	468	-572	3308.13	91.00	2.44	-3.001	0.324	-2.858
5	659	-491	1032.77	6.11	0.32	-0.386	-2.310	-0.346
6	786	-451	1053.00	9.85	0.15	2.365	1.927	-3.036
7	925	-365	294.06	1.87	0.02	0.818	0.367	0.664
8	1056	-335	50.62	0.51	0.01	-2.358	-1.926	-1.929

9	1195	-371	239.50	1.92	0.01	2.856	2.984	2.003
10	1381	-607	260.03	17.23	0.24	-0.830	0.689	-1.425
11	1466	-418	477.75	3.91	0.04	2.719	2.387	-2.824
12	1590	-382	190.98	1.16	0.01	1.223	0.902	1.131
13	1699	-434	62.46	0.42	0.01	-0.656	-1.854	-0.888
14	1845	-390	78.45	0.44	0.00	2.435	2.607	2.526
15	1983	-578	101.14	0.47	0.04	0.739	2.727	0.397
16	2115	-344	52.12	0.36	0.00	2.815	2.932	-2.553
17	2252	-443	70.47	0.46	0.01	1.788	1.928	1.335
18	2375	-508	45.18	0.14	0.02	-0.403	0.672	-0.724
19	2513	-578	82.48	0.44	0.02	2.466	1.294	2.802
20	2628	-960	48.34	1.69	0.04	0.822	-2.103	0.937
21	2752	-476	36.67	0.20	0.01	-0.548	-0.189	-1.058

Table 15 *Formant parameters of the vowels /i/*

Formant number k	Frequency [Hz] f_k	Damping [unit] λ_k	Amplitude1 [unit] A_{1k}	Amplitude2 [unit/s] A_{2k}	Amplitude3 [unit/s2] A_{3k}	Phase1 [rad] φ_{1k}	Phase2 [rad] φ_{2k}	Phase3 [rad] φ_{3k}
1	96	-753	293950	3968.39	16.29	-0.008	-1.927	2.323
2	258	-695	8874.42	268.77	3.40	0.188	-2.751	0.639
3	476	-430	357.74	2.55	0.03	1.716	1.302	2.474
4	580	-404	234.94	1.61	0.01	0.830	0.621	1.377
5	692	-394	135.78	0.68	0.01	0.012	-0.250	-0.115
6	817	-352	23.64	0.20	0.00	2.951	-2.872	-2.826
7	951	-397	199.66	1.38	0.01	1.605	1.896	1.152
8	1070	-546	204.68	2.30	0.10	-0.628	2.756	-0.891
9	1242	-502	171.85	0.73	0.03	2.011	2.354	1.646
10	1422	-628	182.92	6.47	0.11	-1.621	-0.215	-2.301
11	1529	-563	162.86	5.64	0.10	1.546	-0.042	2.061
12	1682	-423	84.11	0.75	0.01	-1.283	-1.955	-0.844
13	1836	-411	51.27	0.25	0.01	2.372	2.720	2.427
14	1955	-441	54.94	0.13	0.01	-0.129	0.654	-0.499
15	2092	-575	65.85	0.58	0.02	2.187	0.704	2.645
16	2237	-410	33.13	0.19	0.00	0.520	0.420	0.246
17	2350	-499	31.86	0.12	0.01	-1.543	-1.371	-1.560
18	2466	-627	40.69	0.95	0.02	2.087	0.763	2.789
19	2700	-514	10.69	0.57	0.02	-1.638	0.304	-2.517

With a help of the obtained parameters, we got 21 quasipolynomial models for the vowel /a/, and 19 quasipolynomial models for the vowel /i/. The plots of

these models are shown in Fig. 25 and Fig. 26.

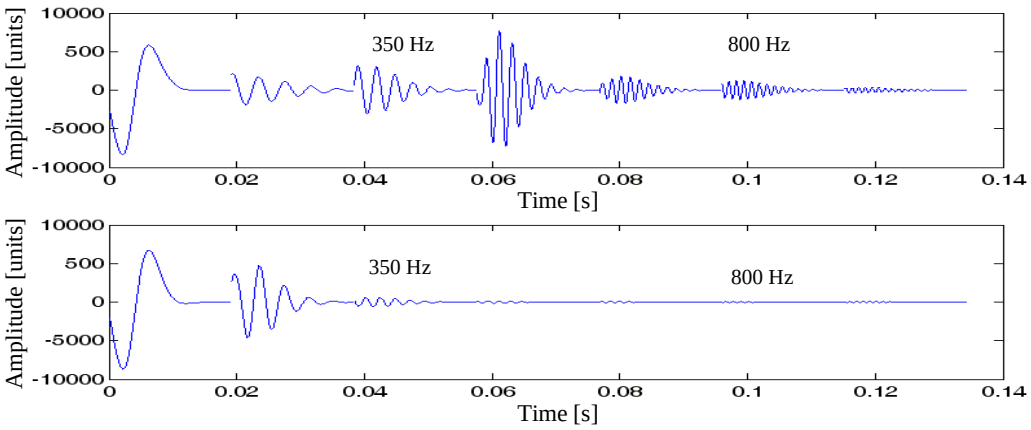


Fig. 25 *The vowel /a/ and /i/ formants with frequencies from the bandwidth of 30-1000 Hz (the upper plot – formants of the vowel /a/, the lower plot - formants of the vowel /i/)*

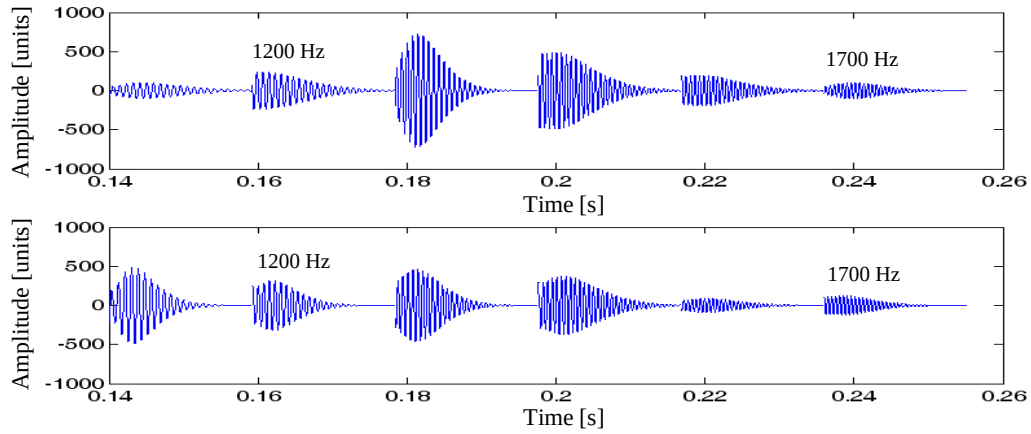


Fig. 26 *The vowel /a/ and /i/ formants with frequencies from the bandwidth of 1001-2000 Hz (the upper plot – formants of the vowel /a/, the lower plot - formants of the vowel /i/)*

After adding the estimated quasipolynomials we obtained the model signal. These models along with the true vowel /a/ and /i/ signals are shown in Fig. 27.

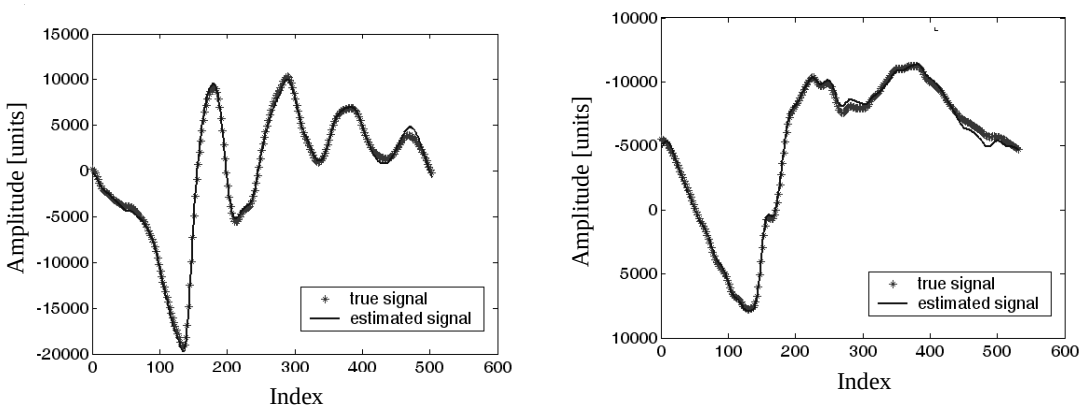


Fig. 27 *The true and estimated signal for the 3-rd formant interval*

The vowel /a/ root-mean-square estimation error is equal to 4.69 %, and this error for the vowel /i/ is equal to 5.22 %.

In order to get the system input impulse amplitudes, the Fourier transform of the diphthong /ai/ was calculated. This transform was then filtered in the intervals shown in Table 13. After filtering, we obtained 21 signals for the vowel /a/ and 19 signals for the vowel /i/ in these intervals. For each of the obtained signals, the inverse Fourier transform is computed and signals in the time domain whose length was 22 560 each is got. Then the local maxima of each of these signals are searched and in columns of two matrices corresponding to the vowels /a/ and /i/ are stored. The values of these columns are shown in Fig. 28.

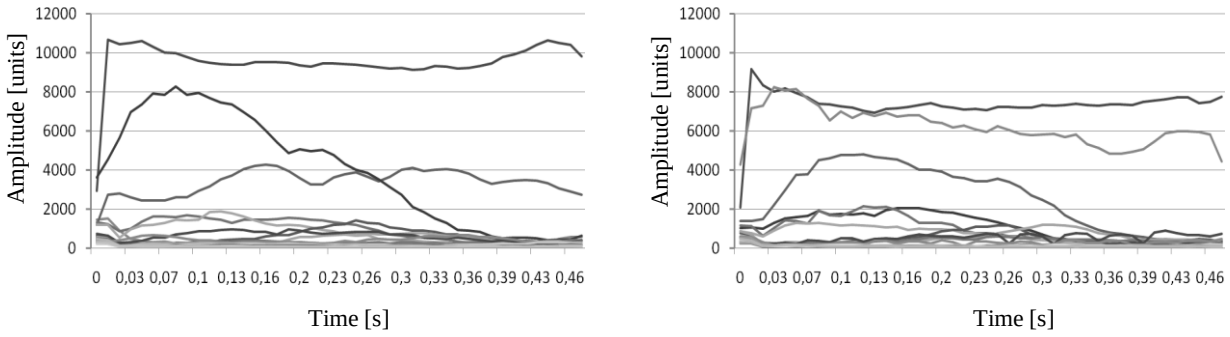


Fig. 28 *The input impulse amplitudes (the left figure – for the vowel /a/, the right figure – for the vowel /i/)*

After analysing the values of these matrices, we see,that these values are rather large. Therefore they are normed.

The normed values are multiplied by values of the tangent and cotangent function. The values corresponding to the vowel /a/ are multiplied by the arccotangent function *arccot(x)* while these corresponding to the vowel /i/ – by the arctangent function (more precisely, by the function *arctan(x)* $+\pi/2$ (see Fig. 29). The result is shown in Fig. 30. Multiplication by these functions are used in order to decrease the input impulse amplitudes for the vowel /i/ in the first half of the diphthong /ai/, and to decrease these amplitudes for the vowel /a/ in the second half of the diphthong /ai/.

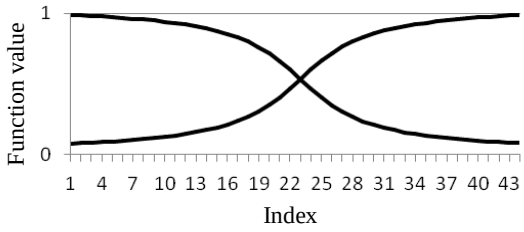


Fig. 29 *The values of the arccotangent function arccot(x) and those of the arctangent function* $\boldsymbol{arctan(x) + \pi/2}$ *used to decrease/increase the input impulse amplitudes for the vowels /a/ and /i/*

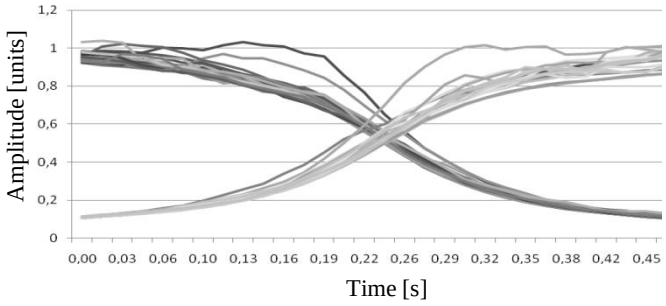
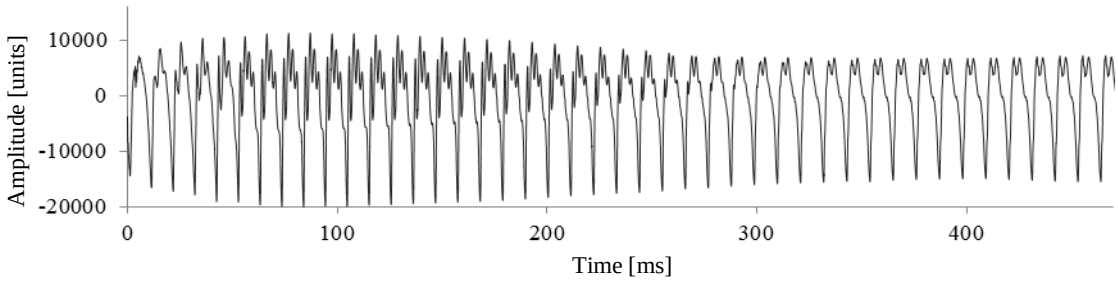
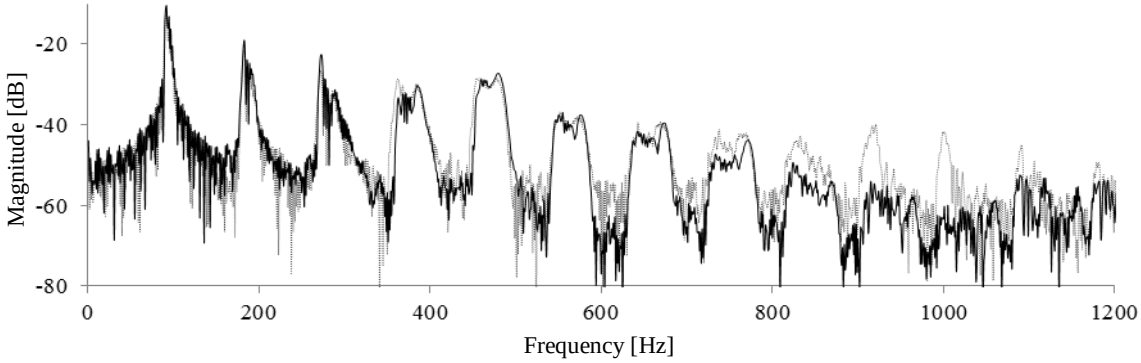


Fig. 30 *Input amplitude dynamics*

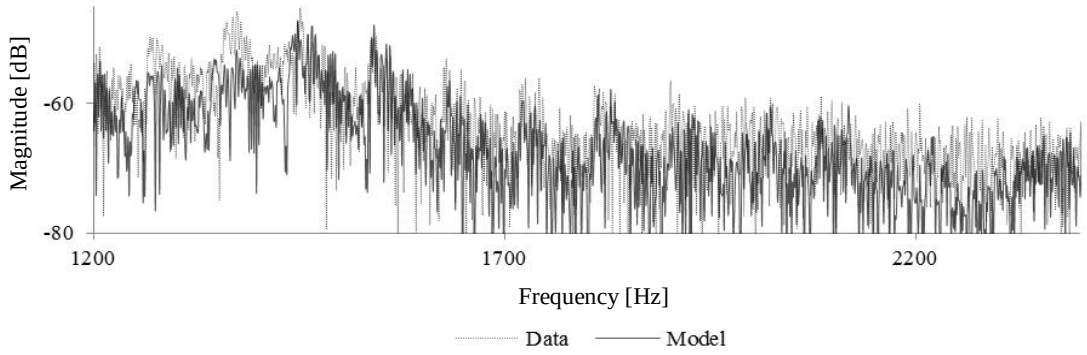
Now, when we have calculated the inputs and impulse response of our MISO system, we obtain the system output using formula (22). In Fig. 31, a fragment of the modelled signal (the diphthong /ai/) and its magnitude response is presented.



a)



b)



c)

Fig. 31 *The Fourier transform of the output process of the synthesyzer "ai": a) the output signal; b) the magnitude response in the range 1-1200 Hz; c) the magnitude response in the range 1200-2400 Hz*

4.4. Joining of vowel and semivowel models

An utterance of the Lithuanian word "laimė" („happiness") were considered. Its duration was 1.34 s. Fig. 32 shows the recorded Lithuanian word "laimė" phonogram:

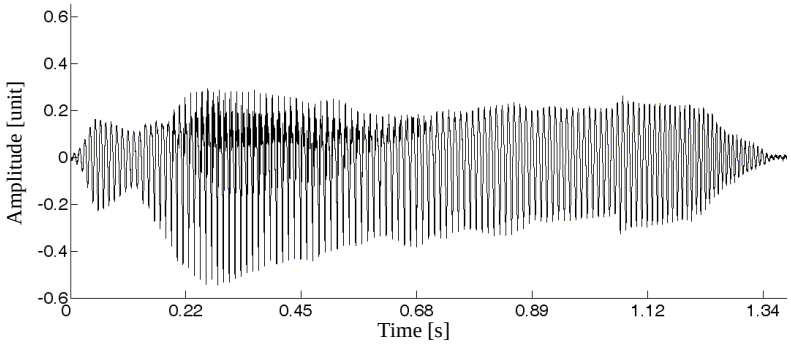


Fig. 32 *The recorded Lithuanian word "laimė"*

A MISO system is used for this word modelling. The input number is equal to the number of phonemes in a word multiplied by the number of formants of the uttered letter. Denote by L the number of phonemes of the synthesized word, and by K_l the number of formants of the uttered *l*-th letter. Each channel is modelled by a single input and single output linear dynamic system having a complex root of multiplicity 3 with an appropriate frequency and damping factor. The impulse response of the channel with the index kl, $0 < k < K_l$, $0 < l < L$ is modelled by a formant that is described by a second degree quasipolynomial:

$$f_{kl}(t) = a_{kl1}e^{-\lambda_{kl}t}\sin(2\pi f_{kl}t + \varphi_{kl1}) + a_{kl2}tf_d e^{-\lambda_{kl}t}\sin(2\pi f_{kl}t + \varphi_{kl2}) + a_{kl3}t^2 f_d^2 e^{-\lambda_{kl}t}\sin(2\pi f_{kl}t + \varphi_{kl3}) \quad (74)$$

where λ_{kl} is the damping factor, f_{kl}- the resonant frequency, a_{kl1}, a_{kl2}, a_{kl3} - amplitudes, φ_{kl1}, φ_{kl2}, φ_{kl3} - phases, f_d - sampling frequency. In (1), $t > 0$ stands for continuous time. Introducing the repeated roots, corresponding to formant quasipolynomials allows, us to get more natural sounds. Each formant of the uttered letter corresponds to a resonance of its magnitude response.

A discrete-time synthesizer with the sampling interval $\Delta t = 1/f_d$ is used. Denote by $h_{kl}(i) = f_{kl}(i \cdot \Delta t)$, $i = 0, 1, 2, ..., N-1$, the unit response of the kl-th channel, by $u_{kl}(n)$ the input sequence of the *k*-th channel of the *l*-th sound. Note

that

$$u_{kl}(n)=\begin{cases}0, & n\neq M_1, M_1+M_2, M_1+M_2+M_3,\ldots\\ x_{kl}(n), & otherwise\end{cases} \tag{75}$$

where $x_{kl}(n)>0$ are real bounded numbers, $M_1, M_2,\ldots$ is the number of samples of the fundamental frequency periods. Then the output $y(n)$ of the synthesized system is represented by the convolution equation:

$$y(n)=\sum_{l=1}^{L}\sum_{k=1}^{K_l}\sum_{i=0}^{N}u_{kl}(n-i)h_{kl}(i) \tag{76}$$

$n=0,1,\ldots.$

A synthesizer scheme is shown in Fig. 33.

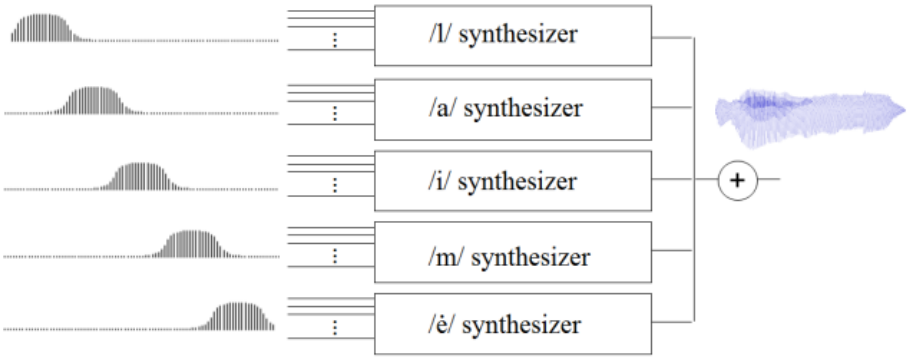


Fig. 33 *The synthesizer scheme of the word "laimė"*

In fact, we use the parallel connection of both the phoneme level and formant level. Excitation signals are inputted successively for each uttered phoneme using overlapping in the transition regions.

Fig. 33 shows that, in order to make the synthesizer, we need to know the formant parameters. Using them, we can get the unit responses of the channels and determine the input sequences. For evaluation of these unknown values, we use the uttered words recorded in computer (see Fig. 32).

After analysing Fig. 32, we can see that both vowels and semivowels have periodic character. In Fig 34, we can see the periodicity of the semivowel /l/.

Fig. 34 *The recorded uttered semivowel /l/*

The dark curve indicates our chosen period using which we create the synthesizer model for the phoneme /l/. The magnitude response of the selected period is calculated, and the obtained magnitude response is divided into 26 frequency bands that are shown in Fig. 35 and Table 16.

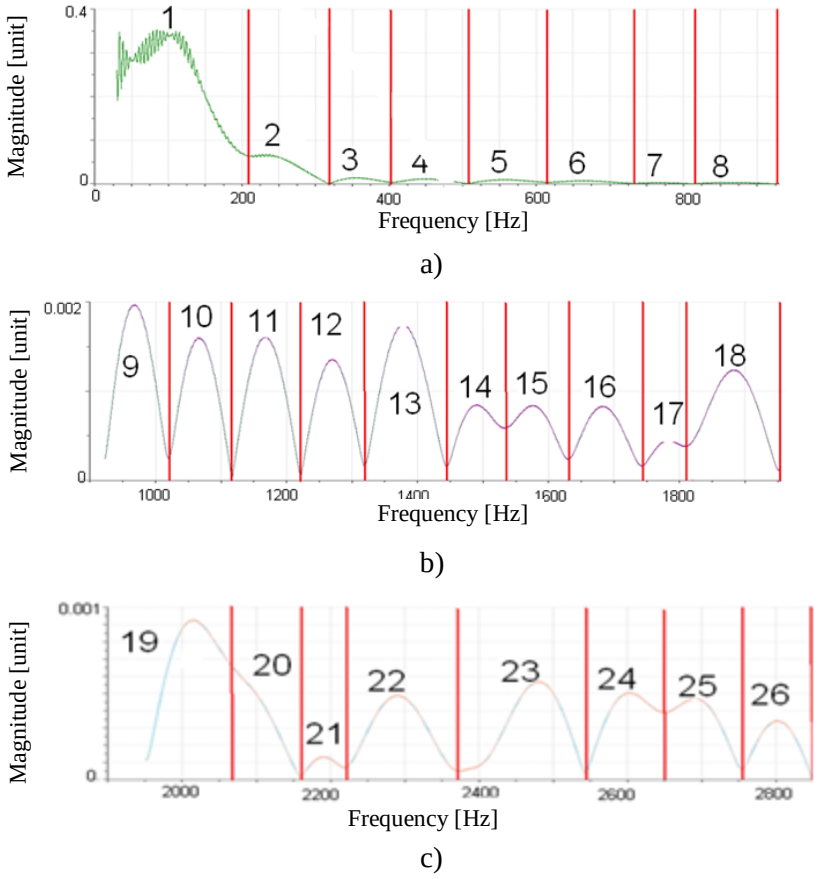


Fig. 35 *The magnitude response of the selected pitch of the semivowel /l/ (the frequency bands: a) 0-922 Hz, b) 923-1950 Hz, c) 1951-2850 Hz)*

Table 16 *The selected frequency bands of the semivowel /l/*

I band	II band	III band
30-210	923-1020	1951-2080
211-315	1021-1117	2081-2160
316-405	1118-1222	2161-2220
406-505	1223-1320	2221-2370
506-615	1321-1445	2371-2545
616-735	1446-1535	2546-2650
736-811	1536-1630	2651-2755
812-922	1631-1745	2756-2850
	1746-1810	
	1811-1950	

For each frequency band, we apply the inverse Fourier transform of the selected pitch and obtain the signals of a simple form that are shown in Fig. 36.

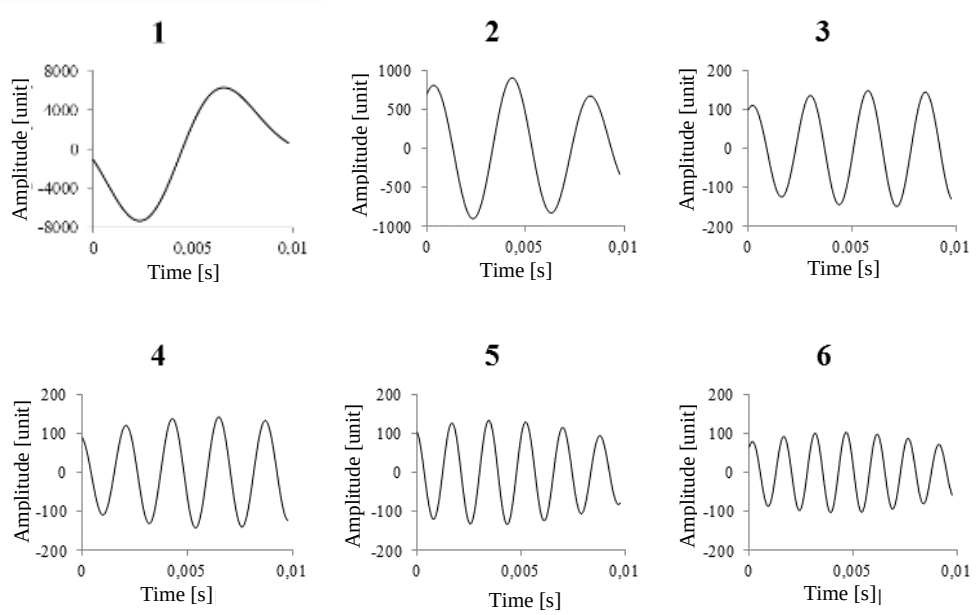


Fig. 36 *The signals of the selected pitch corresponding to the frequency bands 1-6*

Each of the obtained signals is modelled by formula (74). The parameters are estimated by parameter estimation algorithm described in Section 3.4. It is also assumed that the filter impulse response decays after three periods, thus we use the convoluted basis signal matrix $\mathbf{\Phi}$ defined by (45). The parameters of the first six formants of the semivowel /l/ are shown in Table 17.

Table 17 *The formant parameters for the semivowel /l/*

f_{k1}	λ	a_{k11}	a_{k12}	a_{k13}	φ_{k11}	φ_{k12}	φ_{k13}
104	-676	109266	1649	7,8	0	-1,9	2,3
249	-304	1329	7	0,04	0,9	0,9	1,1
362	-377	235	0,52	0,02	1,1	-3,0	0,9
451	-309	223	0,55	0,01	1,8	1,2	2,0
555	-307	196	1,07	0,01	1,9	1,9	2,3
674	-308	146	0,82	0,01	1,1	1,0	0,4

After calculating the parameters, the error is evaluated. Three unit impulses are input and we get the pitch model that we compare with the real signal pitch. The true and estimated signal pitches are shown in Fig. 37.

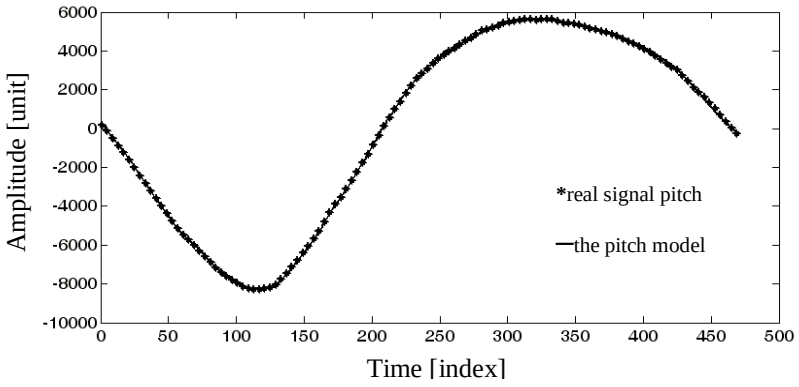


Fig. 37 *The true and estimated signal pitches*

The root-mean-square error (RMSE) for the semivowel /l/ is equal to 0.90 %. For every phoneme, its characteristic pitch is selected and its parameters are calculated. The parameter estimation accuracy is shown in Table 18.

Table 18 *The formant parameter estimation error*

Phoneme	**RMSE**
/l/	0.90 %
/a/	4.69 %
/i/	3.84 %
/m/	1.47 %
/ė/	0.66 %
/e/	2.00 %
/ę/	2.98 %
/j/	0.46 %

The impulse responses of the phoneme formants described as second-degree quasipolynomial models are calculated in accordance with the formula (74). Fig. 38 shows the sums of formants of the estimated impulse responses for each phoneme.

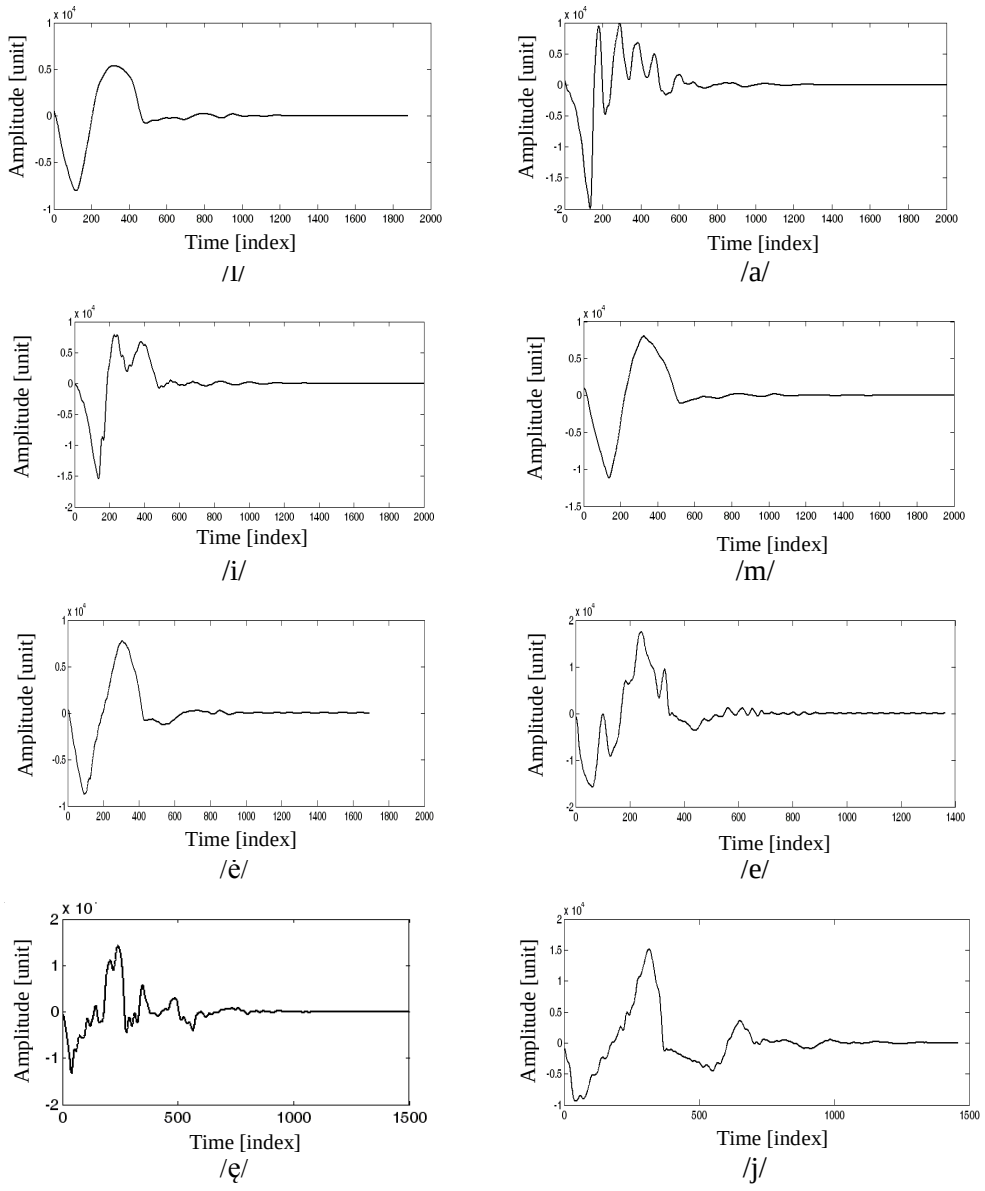


Fig. 38 *The sums of formants of the estimated impulse responses for each phoneme*

Fig. 38 reveals that the vowel impulse response oscillation is stronger than those of the semivowels.

Now a problem arises: what inputs have to be given. First of all, the fundamental frequency of the word "laimė" is estimated. For this goal, the signal of the word "laimė" is filtered using a rectangle narrowband filter (80-140 Hz). The filtering is done with the help of the inverse Fourier transform. After obtaining the signal zeroes, the durations of the fundamental frequency periods are calculated. The fundamental frequency is calculated by the following formula:

$$f_i = \frac{1}{T_i}, \tag{77}$$

where T_i is the length of the i-th pitch.

The trajectories of the fundamental frequency of all the cases of the word "laimė" are shown in Fig. 39.

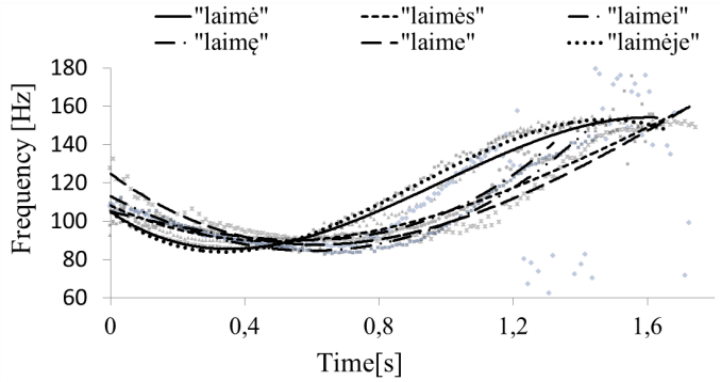


Fig. 39 *The trajectories of the fundamental frequency of the word "laimė" and its cases*

The trend formulas for the fundamental frequency of the word "laimė" and its cases are presented below:

$$f_1(t) = f_0 \cdot (-0.9t^3 + 2.4t^2 - 1.3\mathrm{t} + 1) + \varepsilon(\mathrm{t}) \tag{78}$$

where f_0 – the fundamental frequency (in our case $f_0 = 105\,Hz$); $\mathrm{t} = 0,\ T_1,\ T_2, \ldots$; $\{\varepsilon(\mathrm{t})\}$ – independent Gaussian random variables with zero mean and variance $\sigma^2 = 4$.

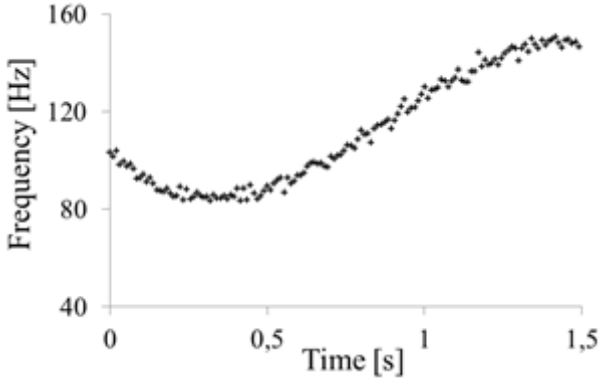


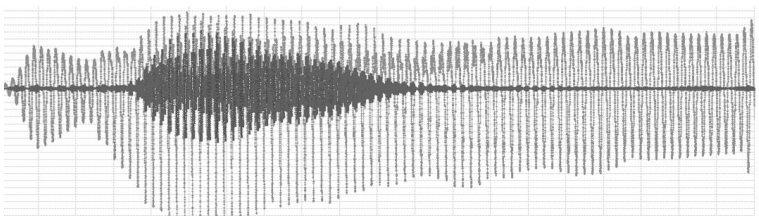
Fig. 40 *The trajectories of the fundamental frequency of the word "laimė" synthesizer*

The simplest synthesizer excitation scenario is to give unit impulses into all the inputs consequently, i. e. in the beginning, into the phoneme /l/ synthesizer, then into the phoneme /a/ synthesizer, and so on. Our investigation, however, revealed that the phoneme pitches are of different length and the pitch amplitudes differ. Moreover, there exist transitions between phonemes.

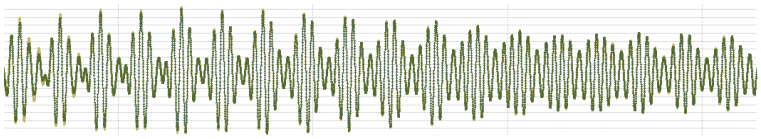
The fundamental frequency can help to determine the time instants when it is necessary to give inputs. We need to know what input must be given for each formant of each phoneme.

For this purpose the recorded signal is filtered by rectangular filters whose bands are specified by formants of the investigated phoneme. The output signals of these filters are divided into periods.

An example of signal filtering with a narrowband filter is presented in Fig. 41.



a)



b)

Fig. 41 *Extraction of the 4th formant of the phoneme /a/: a) the word "laimė" and the 4th formant (dark part), b) the 4-th formant in enlarged time scale*

After filtering the whole sound in the frequency bands corresponding to phonemes and calculating the maximums, we obtain the input signals. The maximum values of the word "laimė" received when splitting the spectrum according to the formants of the phoneme /a/ are shown in Fig.42.

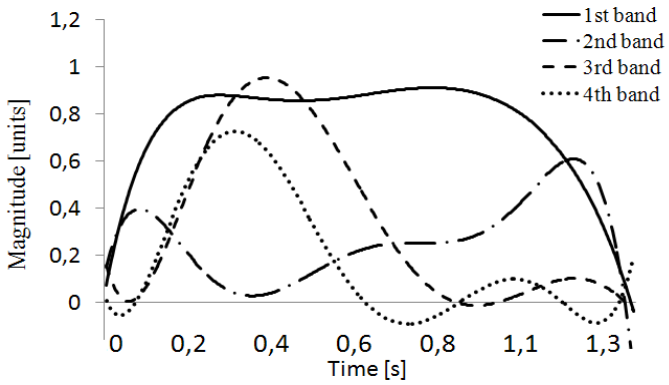


Fig. 42 *Filtering results of the word "laimė" in the frequency bands corresponding to the phoneme /a/ formants*

In the same way, the sound signal „laimė” using rectangular filters is filtered in the bands corresponding to formants, and transitions are created between sounds using the parabola curve. As an example, parabola curve for input calculation of the phoneme /ė/ is applied. Fig. 43 shows in the sequence of phonemes input.

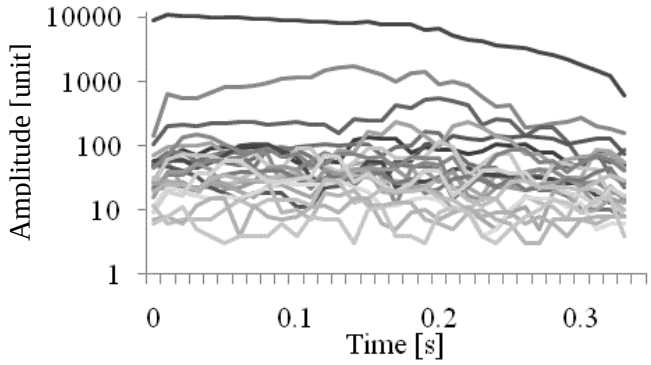


Fig. 43 *The input values of the phoneme /ė/*

The obtained input values are normalized and their trends are calculated. The trends of the values of the first four inputs are presented in Fig. 44.

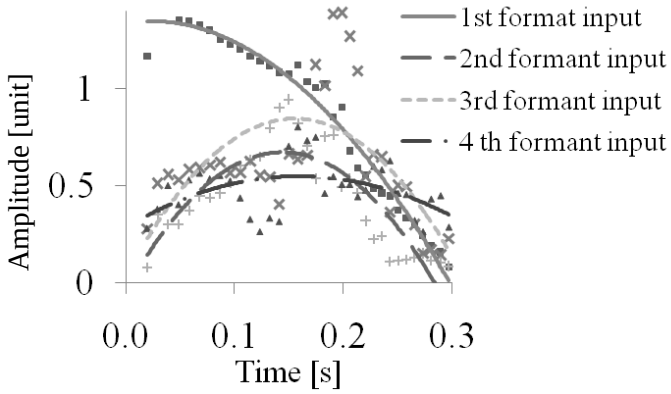


Fig. 44 *The trends of the input values of the 1-4 formant of the phoneme /ė/*

In order to describe the inputs by formulas, we need to determine the parameters y_{max}, t_{max} and t_{lim} (see Table 19).

Table 19 *The parabola parameters of the 1-4 formants of the phoneme /ė/*

1st formant	y_{max}	1.9
	t_{max}	0.02
	t_{lim}	1
2nd formant	y_{max}	1
	t_{max}	0
	t_{lim}	1

3rd formant	y_{max}	1.4
	t_{max}	0.17
	t_{lim}	1
4th formant	y_{max}	1.1
	t_{max}	0.08
	t_{lim}	0.5

Using the parameters shown in Table 19, we obtain the parabolas describing the input amplitudes.

Fig. 45 shows how each phoneme is excited.

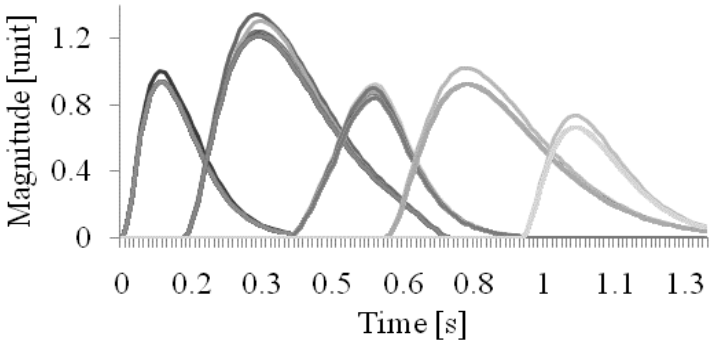
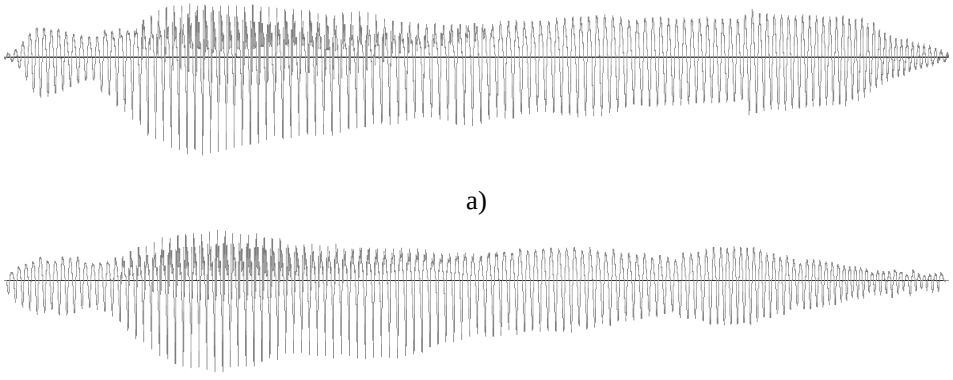


Fig. 45 *The inputs of the word "laimė" synthesizer*

The true and synthesized word "laimė" are shown in Fig. 46.



a)

b)

Fig. 46 *The word "laimė": a) true, b) synthesized*

The magnitude response of the true and synthesized word "laimė" is shown in Fig. 47.

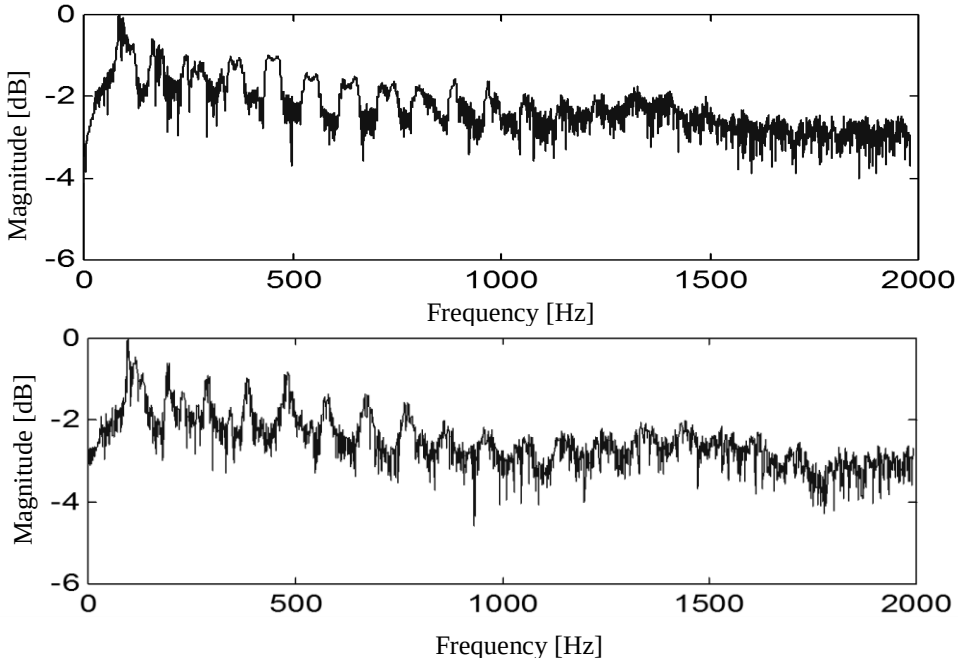


Fig. 47 *The magnitude response of the true (the upper plot) and synthesized (the lower plot) word "laimė"*

4.5. Conclusions of Section 4

Research has shown that the estimates of the fundamental frequency obtained by the MUSIC method are less scattered around their average if compared with the ones obtained by the DFT method.

Approximation of a real sound signal by a sum of the first 10 harmonics with the fundamental frequency obtained by the MUSIC and DFT methods gave a smaller error in the case of the MUSIC method.

The harmonic method uses a higher-order model with a larger number of parameters in comparison with the formant method, but the sounds synthesized by the harmonic method sound more naturally.

The harmonic synthesis method is recommended to be used in speech synthesis by computers meanwhile the formant synthesis method - in devices where the memory size is limited as in mobile phones.

The small estimation errors and audio test show that the proposed framework gives sufficiently good vowel and semivowel synthesis quality.

Calculations have shown that the parameter estimation method is stable for all the considered sounds.

The accuracy of diphthong modelling was high. It was almost impossible to distinguish between real and simulated diphthongs in various Lithuanian words with a help of audiotesting. Only the magnitude response of the whole signal of the simulated diphthong differed a little from the magnitude response of the recorded data in some frequency regions.

During the simulation of input sequences, special attention was paid to the transitions between phonemes. The input sequence of each phoneme has been described by three parabolas.

Simulation has revealed that the word consisting of vowels and semivowels obtained with the proposed synthesis method is good enough and it is difficult to distinguish it from the real one. The quality of the synthesized sound was significantly improved due to input transitions.

The method proposed can be applied to the words composed of vowels and semivowels.

5

Conclusions

The research object of the dissertation is Lithuanian vowel and semivowel phoneme models. In order to develop models for vowels and semivowels, the main characteristics of these sounds have been identified. A phoneme synthesis framework that is based on a vowel and semivowel phoneme mathematical model and an automatic procedure of estimation of the phoneme fundamental frequency and input determining has been proposed. Within this framework two synthesis methods have been given: the harmonic method and formant method.

The research completed in this thesis has led to the following conclusions:

1. Lithuanian language has ninety two phonemes. Twenty eight of them are pure vowel phonemes, nineteen – semivowel phonemes. In general case, the character of vowel and semivowel signals is periodic.
2. The fundamental frequencies of the stressed vowels and semivowels are lower than those of the unstressed ones.

3. The estimates of the fundamental frequency obtained by the MUSIC method are less scattered around their average if compared with the ones obtained by the DFT method. The methods for Lithuanian male vowels /a/, /i/, /o/, /u / were applied. The smallest standard deviation 2.36 Hz was obtained by the MUSIC method for the vowel /i/, and the largest – 5.59 Hz – by the DFT method for the vowel /a/.
4. The harmonic method uses a higher-order model with a larger number of parameters in comparison with the formant method but the sounds synthesized by the harmonic method sound more naturally. The average RMSE for the estimated signal spectrum for all the male and female vowels is equal to 13.9 % in the formant method case and 12.4 % in the harmonic method case. The average RMSE for the estimated signal spectrum for all the male and female semivowels is equal to 19.9 % in the formant method case and 16.7 % in the harmonic method case.
5. The average time of the phoneme parameter estimation for all the male and female vowels and semivowels is equal to 16.1 s in the formant method case and 37.2 s in the harmonic method case. The average time of the phoneme synthesis for all the male and female vowels and semivowels is equal to 0.09 s in the formant method case and 0.44 s in the harmonic method case.
6. The accuracy of diphthong modelling is high. It is almost impossible to distinguish between the real and simulated diphthongs in various Lithuanian words with a help of audiotesting. Only the magnitude response of the whole signal of the simulated diphthong differs a little from the magnitude response of the recorded data in some frequency bands.

7. Simulation has revealed that the word consisting of vowels and semivowels obtained with the proposed synthesis method is good enough and it is difficult to distinguish it from the real one. The quality of the synthesized sound was significantly improved due to input transitions.
8. The created automatic system will suit any speaker, any vowel and semivowel phoneme.

References

Anbinderis, T. (2010). Mathematical modelling of some aspects of stressing a Lithuanian text. Doctoral dissertation, Vilnius University, Vilnius [in Lithuanian].

Balbonas, D., Daunys, G. (2007). Movement of Formants of Vowels in Lithuanian Language. Electronics and Electrical Engineering, 7(79), 15–18.

Balbonas, D. (2009). Analysis of Vowels Spectrum. Doctoral dissertation. Kaunas University of Technology, Kaunas.

Bastys, A., Kisel, A., Šalna, B. (2010). The Use of Group Delay Features of Linear Prediction Model for Speaker Recognition. Informatica, 1(21), 1–12.

Borzone De Manrique, A. M. (1979). Acoustic Analysis of the Spanish Diphthongs. Phonetica, 3(36), 194 – 206.

Burk, P. (2005). Music and Computers Course Guide: A Theoretical & Historical Approach, Key College Publishing.

de Cheveigné, A., Kawahara. H. (2002). YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 111(4), 1917-1930.

Carlson R., Sigvardson T., Sjölander A. (2002). Data-driven formant synthesis. Proceedings of Fonetik, TMH-QPSR, 44 (1), 121-124.

Collins Concise English Dictionary, 2009.
Accessed at: http://www.thefreedictionary.com/diphthong

Cook, P. R. (2002). Real Sound Synthesis for Interactive Application. A K Peters Ltd., 263 p.

Cressey, W. W. (1978). Spanish Phonology and Morphology. A Generative View. Georgetown University Press, 169 p.

Čeidaitė, G., Telksnys, L. (2010).Analysis of Factors Influencing Accuracy of Speech Recognition. Electronics and Electrical Engineering, 9(105), 69–72.

Donovan, R. E. (1996). Trainable Speech Synthesis. Doctoral dissertation. Cambridge University, Cambridge.

Driaunys, K., Rudžionis, V. E., Žvinys, P. (2009). Implementation of hierarchical phoneme classification approach on LTDIGITS corpora. Information technology and control, 4(38), 303-310.

Encyclopedia Britannica. Encyclopedia Britannica Online. Encyclopedia Britannica Inc., 2012

Fant, G. (1970). Acoustic Theory of Speech Production. Mouton & Co, 328 p.

Fox, A. (2005). The Structure of German. Oxford Linguistics, 334 p.

Frolov, A., Frolov, G. (2003). Speech Synthesis and Recognition. Modern Solutions (eBook). Accessed at: http://www.frolov-lib.ru/books/hi/index.html

Golub, G., Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM Journal or Numerical Analysis, 2(10), 413 – 432.

Garšva, K. (2001). Complex diphthongs ie, uo and their phonological interpretation. Žmogus ir žodis (A man and a word). Vilnius Pedagogical University. Didaktinė lingvistika (Didactic linguistics). Mokslo darbai (Research works). Edited by V. Drotvinas, 3(1), 23–26 [in Lithuanian].

Geumann, A. (1997). Formant trajectory dynamics in Swabian diphthongs. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, vol. 35, 35 – 38.

Girdenis, A. (1995). Teoriniai fonologijos pagrindai (Theoretical basics of phonology). Vilnius University, Vilnius [in Lithuanian].

Handbook of the International Phonetic Association, Cambridge University Press, 1999.

Hess, W. (1983). Pitch Determination of Speech Signals, Springer-Verlag, New York.

Holmes J., Holmes W. (2001). Speech Synthesis and Recognition,. CRC Press, 298 p.

Hopcroft, J. E., Ullman J. D. (1979). Introduction to Automata Theory, Languages and Computation, Addison-Wesley Publishing, Reading Massachusetts.

Ivanovas E., Navakauskas D. (2010). Development of biometric systems for person recognition: Biometric feature systems, traits and acquisition. Electronics and Electrical Engineering, 5(101), 87-90.

Ivanovas, E., Navakauskas, D. (2011). Peculiarities of Wiener Class Systems and their Exploitation for Speech Signal Prediction. Electronics and Electrical Engineering, 5(111), 107–110.

Ivanovas E., (2012). Development and implementation of means for word duration signal processing (doctoral dissertation), Vilnius: Technika, 128 p. [in Lithuanian].

Janicki, A. (2004). Selected Methods of Quality Improvement in Concatenative Speech Synthesis for the Polish Language. Doctoral dissertation. Warsaw University of Technology, Warsaw.

Kajackas, A., Anskaitis, A. (2009). An Investigation of the Perceptual Value of Voice Frames. Informatica, 4(20), 487–498.

Kamarauskas, J. (2009). Speaker recognition by voice (doctoral dissertation), Vilnius: Technika, 124 p. [in Lithuanian].

Kasparaitis, P. (1999). Transcribing of the Lithuanian text using formal rules. Informatica, 10(4), 367–376.

Kasparaitis, P. (2000). Automatic stressing of the Lithuanian text on the basis of a dictionary. Informatica, 11(1), 19–40.

Kasparaitis, P. (2001). Text-to-Speech Synthesis of Lithuanian Language. Doctoral dissertation, Vilnius University, Vilnius [in Lithuanian].

Kasparaitis, P. (2005). Diphone databases for Lithuanian text-to-speech synthesis. Informatica, 2(16), 193– 202.

Kasparaitis, P. (2008). Lithuanian Speech Recognition Using the English Recognizer. Informatica, 4(19), 505–516.

Kazlauskas, K. (1999). Noisy Speech Intelligibility Enhancement. Informatica, 2(10), 171–188.

Kazlauskas, K., Pupeikis, R. (2013**).** On intelligent extraction of an internal signal in a Wiener System consisting of a linear block followed by hard-nonlinearity. 1(24), 35-58.

Laurinčiukaitė, S., Lipeika, A. (2007). Framework for Choosing a Set of Syllables and Phonemes for Lithuanian Speech Recognition. Informatica, 3(18), 395–406.

Levenberg, K. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. The Quarterly of Applied Mathematics 2, pp. 164–168.

Lileikytė, R., Telksnys, L. (2011). Quality Estimation Methodology of Speech Recognition Features. Electronics and Electrical Engineering, 4(110), 113–116.

Lileikytė. R., Telksnys, L. (2012). Quality Measurement of Speech Recognition Features in Context of Nearest Neighbour Classifier. Electronics and Electrical Engineering. – Kaunas: Technologija, 2(118), 9–12.

Lipeikienė, J., Lipeika, A. (1998). Language engineering in Lithuania. Informatica, 9(4), 449–456.

Lipeika, A., Lipeikienė, J., Telksnys, L. (2002). Development of Isolated Word Speech Recognition System. Informatica, 1(13), 37–46.

Lipeika, A., Lipeikienė, J. (2003). Word Endpoint Detection Using Dynamic Programming. Informatica, 4(14), 487–496.

Lipeika, A., Lipeikienė, J. (2008). On the Use of the Formant Features in the Dynamic Time Warping Based Recognition of Isolated Words. Informatica, 2(19), 213–226.

Lipeika, A. (2010). Optimization of Formant Feature Based Speech Recognition. Informatica, 3(21), 361–374.

Lithuanian speech synthesis web page (2010). Accessed at: http://www.garsiai.lt

Mannell, R. H. (1998). Formant diphone parameter extraction utilising a labelled single-speaker database. Proceedings of the Fifth International Conference on Spoken Language Processing. Sydney, Australia, 30 November – 1 December, 1998.
Accessed at: http://clas.mq.edu.au/rmannell/research/icslp98/iclsp98_mannell.pdf

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. SIAM Journal on Applied Mathematics 11, 431–441.

Markel, J. D., Gray, A. H. (1976). Linear Prediction of Speech. Springer Verlag, Berlin.

Martirosian, O., Davel, M. (2008). Acoustic analysis of diphthongs in Standard South African English. Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa, Cape Town, South Africa, 27-28 November, 153-157

Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V. (2009). Investigation of foreign languages models for Lithuanian speech recognition. Electronics and Electrical Engineering, No. 3, 15-20.

Maskeliūnas, R., Ratkevicius, K., Rudzionis, V. (2011). Voice-based Human-Machine Interaction Modeling for Automated Information Services. Electronics and Electrical Engineering, 4(110), 109–112.

Mažonavičiūtė, I., Baušys, R. (2011). Translingual Visemes Mapping for Lithuanian Speech Animation. Electronics and Electrical Engineering, 5(111), 95–98.

Milivojevic, Z., Mirkovic, M., Milivojevic, S. (2006). An estimate of fundamental frequency using PCC interpolation – comparative analysis. Information Technology and Control, 35(2), 131-136.

Mobius, B., van Santen, J. (1996). Modeling segmental duration in German text-to-speech synthesis. Proceedings of the Fourth International Conference on Spoken Language Processing. Philadelphia, USA, October 3-6, 1996, vol. 4, 2395-2398

Accessed at: http://www.asel.udel.edu/icslp/cdrom/vol4/652/a652.pdf

Murakami T., Ishida Y. (2001). Fundamental frequency estimation of speech signals using MUSIC algorithm. Acoust. Sci. Technol, 22(4), 293 – 297.

Norkevičius, G., Raškinis, G. (2008). Modeling Phone Duration of Lithuanian by Classification and Regression Trees, using Very Large Speech Corpus. Informatica, 2(19), 271–284.

Navakauskas, D., Paulikas, Š. (2006). Autonomous robot in the adverse environment: Intelligent control by voice. Solid State Phenomena, vol. 113 Mechatronic Systems and Materials, 325-329.

Onelook, 2010. Accessed at: http://www.onelook.com/?w=phone&ls=a

Oxford dictionaries, 2010.
Accessed at: http://www.askoxford.com/concise_oed/phoneme?view=uk

Raškinis, G., Raškinienė, D. (2003). Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. Informatica, 14(1), 75–84.

Ringys, T., Slivinskas, V. (2009). Formant modelling of Lithuanian language vowel natural sounding. The Materials of the 4th International Conference on Electrical and Control Technologies ECT-2009. Kaunas: Technologija, 5-8 [in Lithuanian].

Ringys, T., Slivinskas, V. (2010). Lithuanian language vowel formant modelling using multiple input and single output linear dynamic system with multiple poles. Proceedings of the 5th International Conference on Electrical and Control Technologies ECT-2010. 6-7 May, Kaunas, Lithuania, 117-120.

Rudžionis, A., Ratkevičius, K., Dumbliauskas, T., Rudžionis, V. (2008). Control of computer and electric devices by voice. Electronics and Electrical Engineering, No. 6, 11-16.

SIL International, 2004. Accessed at:
http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsAPhone.htm

Šilingas, D., Telksnys, L. (2004). Specifics of Hidden Markov Model Modifications for Large Vocabulary Continuous Speech Recognition. Informatica, 1(15), 93–110.

Šimonytė, V., Slivinskas, V. (1997). Estimation of multiple exponential-sinusoidal models. Theory of Stochastic Processes, 3 – 4(19), 426 – 435.

Schmidt, R.O. (1986). Multiple Emitter Location and Signal Parameter Estimation, IEEE Trans. Antennas Propagation, 34(3), 276-280.

Skripkauskas, M., Telksnys, L. (2006). Automatic Transcription of Lithuanian Text Using Dictionary. Informatica, 4(17), 587–600.

Slivinskas, V., Šimonytė, V. (1990). Minimal realization and formant analysis of dynamic systems and signals. Mokslas. Vilnius 168 p. [in Russian] (republished by Booksurge, USA, 2007).

Slivinskas, V., Šimonytė, V. (1997). Estimation of quasipolynomials in noise: theoretical, algorithmic and implementation aspects. Computer-Intensive Methods in Control and Signal Processing: The Curse of Dimensionality. Birkhauser, Boston, 223 – 235.

Stoica, P., Moses, R. (1997). Introduction to Spectral Analysis. Englewood Cliffs, Prentice-Hall.

Stoica, P., Nehorai, A. (1989). MUSIC, Maximum Likelihood, and Cramer-Rao Bound, IEEE Trans. Acoustics, Speech, and Signal Processing, 37(5), 720-741.

Šalna, B., KamarauskasJ. (2010). Evaluation of Effectiveness of Different Methods in Speaker Recognition. Electronics and Electrical Engineering, 2(98), 67–70.

Šveikauskienė, D. (2005). Graph Representation of the Syntactic Structure of the Lithuanian Sentence. Informatica, 3(16), 407–418.

Tamulevičius, G. (2008). Development of isolated word recognition systems (doctoral dissertation), Vilnius: Technika, 124 p. [in Lithuanian].

Tamulevičius, G., Arminas, V., Ivanovas, E., Navakauskas, D. (2010). Hardware accelerated FPGA implementation of Lithuanian isolated word recognition system. Electronics and Electrical Engineering, 3(99), 57-62.

The Linguistics Research Center (LRC), 2013
Accessed at: www.utexas.edu/cola/centers/lrc/eieol/litol-2-R.html

The MBROLA Project. [http://tcts.fpms.ac.be/synthesis]

Therrien, C. W. (1992). Discrete Random Signals and Statistical Signal Processing. Englewood Cliffs, Prentice-Hall, 614–655.

List of Publications

Articles in the reviewed scientific periodical publications

A 1. G. Pyž, V. Šimonytė, V. Slivinskas (2011). Modelling of Lithuanian Speech Diphthongs. *Informatica*, Vol. 22 (3), p. 411-434. ISSN 0868-4952 [ISI Web of Science].

A 2. G. Pyž, V. Šimonytė, V. Slivinskas (2011). Joining of Vowel and Semivowel Models in Lithuanian Speech Formant-based Synthesizer. *Proc. of the 6th International Conference on ECT-2011*, p. 114-119, ISSN 1822-5934 [ISI Proceedings].

A 3. G. Pyž, V. Šimonytė, V. Slivinskas (2012). Lithuanian Speech Synthesizing by Computer Using Additive Synthesis. *Elektronika ir elektrotechnika*, Vol. 18 (8), p. 77-80. ISSN 1392-1215 [ISI Web of Science].

A 4. V. Šimonytė, G. Pyž, V. Slivinskas (2009). Application of the MUSIC method for estimation of the signal fundamental

frequency. *Lietuvos matematikos rinkinys. LMD darbai*, T. 50, p. 391-396, ISSN 0132-2818.

A 5. G. Pyž, V. Šimonytė, V. Slivinskas (2012). An automatic system of Lithuanian speech formant synthesizer parameter estimation. *Proc. of the 7th International Conference ECT-2012*, p. 36-39, ISSN 1822-5934.

A 6. V. Slivinskas, V. Šimonytė, G. Pyž (2013). Control of Computer Programs by Voice Commands. *Proc. of the 8th International Conference ECT-2013*, p. 37-40, ISSN 1822-5934.

Methodical work

V. Šimonytė, G. Pyž, V. Slivinskas (2010). Signals and their parameter estimation, Vilnius: BMK, ISBN 978-9955-88-44-4 [in Lithuanian].

Appendices

Appendix A. The Lithuanian phoneme list along with the examples

Table 1 *A list of Lithuanian phonemes (originally created by A. Girdenis (Girdenis, 1995) and appended by P. Kasparaitis (Kasparaitis, 2005)) with the examples of their usage*

	Phoneme	***Description***	***Example***
1.	/_/	Pause	-
2.	/a/	The short unstressed vowel /a/	m**a**mà 'mother'
3.	/a`/	The short stressed vowel/a/	lazd**à** 'stick'
4.	/a:/	The long unstressed vowel /a/	dr**ą**sà 'courage'
5.	/a:´/	The long vowel /a/ stressed with the falling accent	k**á**rdas 'sword'
6.	/a:~/	The long vowel /a/ stressed with the rising accent	**ã**čiū 'thank you'
7.	/e/	The short unstressed vowel /e/	m**e**dãlis 'medal'
8.	/e`/	The short stressed vowel /e/	sug**è**sti 'turn bad', 'get out of order'
9.	/e:/	The long unstressed vowel /e/	gr**ę**žinỹs 'well', 'borehole'
10.	/e:´/	The long vowel /e/ stressed with the falling accent	**é**rkė 'mite'
11.	/e:~/	The long vowel /e/ stressed with the rising accent	gyv**ẽ**nimas 'life'
12.	/ė:/	The long unstressed vowel /ė/	k**ė**dẽ 'chair'
13.	/ė:´/	The long vowel /ė/ stressed with the falling accent	up**é**takis 'trout'
14.	/ė:~/	The long vowel /ė/ stressed with the rising accent	gėl**ẽ** 'flower'
15.	/i/	The short unstressed vowel /i/	l**i**gà 'disease', 'illness'
16.	/i`/	The short stressed vowel /i/	k**ì**škis 'rabbit'
17.	/i:/	The long unstressed vowel /i/	t**y**là'silence', 'quiet'

18.	/i:´/	The long vowel /i/ stressed with the falling accent	r**ý**tas 'morning'
19.	/i:˜/	The long vowel /i/ stressed with the rising accent	arkl**ỹ**s 'horse'
20.	/ie/	The gliding unstressed diphthong /ie/	p**ie**štùkas 'pencil'
21.	/i´e/	The gliding diphthong /ie/ stressed with the falling accent	**íe**tis 'spear', 'javelin'
22.	/ie˜/	The gliding diphthong /ie/ stressed with the rising accent	p**iẽ**tūs 'lunch', 'south'
23.	/o/	The short unstressed vowel /o/	**o**žkà 'she-goat'
24.	/o`/	The short stressed vowel /o/	ch**ò**ras 'choir'
25.	/o:/	The long unstressed vowel /o/	k**o**vótojas 'fighter'
26.	/o:´/	The long vowel /o/ stressed with the falling accent	š**ó**nas 'side'
27.	/o:˜/	The long vowel /o/ stressed with the rising accent	Ad**õ**mas 'Adam'
28.	/u/	The short unstressed vowel /u/	k**u**ltūrà 'culture'
29.	/u`/	The short stressed vowel /u/	**ù**pė 'river'
30.	/u:/	The long unstressed vowel /u/	k**ū**rinỹs 'work', 'piece'
31.	/u:´/	The long vowel /u/ stressed with the falling accent	l**ū́**pa 'lip'
32.	/u:˜/	The long vowel /u/ stressed with the rising accent	m**ū̃**šis 'battle'
33.	/uo/	The gliding unstressed diphthong /uo/	**uo**giẽnė 'jam'
34.	/u´o/	The gliding diphthong /uo/ stressed with the falling accent	j**úo**das 'black'
35.	/uo˜/	The gliding diphthong /uo/ stressed with the rising accent	ag**uõ**na 'poppy'
36.	/b/	The consonant /b/	**b**rólis 'brother'
37.	/b"/	The soft (palatalised) consonant /b/	la**b**iáu 'more'
38.	/d/	The consonant /d/	**d**árbas 'work'
39.	/d"/	The soft consonant /d/	liū**d**esỹs 'sadness'
40.	/g/	The consonant /g/	**g**añdras 'stork'
41.	/g"/	The soft consonant /g/	**g**érvė 'crane'
42.	/k/	The consonant /k/	**k**ãtinas 'cat'
43.	/k"/	The soft consonant /k/	**k**iaũlė 'pig'
44.	/p/	The consonant /p/	**p**óvas 'peacock'
45.	/p"/	The soft consonant /p/	**p**etelìškė 'butterfly'
46.	/t/	The consonant /t/	**t**ãkas 'path'
47.	/t"/	The soft consonant /t/	šal**t**ẽkšnis 'black alder'
48.	/c/	The consonant /c/	**c**ãras 'tsar'
49.	/c"/	The soft consonant /c/	**c**itrinà 'lemon'
50.	/č/	The consonant /č/	**č**árdašas 'czardas'
51.	/č"/	The soft consonant /č/	**č**iužinỹs 'mattress'
52.	/dz/	The consonant /dz/	**Dz**ūkijà 'Dzukija' (a region of Lithuania)
53.	/dz"/	The soft consonant /dz/	**dz**ìngtelėti 'clink'
54.	/dž/	The consonant /dž/	**dž**áulis 'Joule'
55.	/dž"/	The soft consonant /dž/	**dž**iaũgsmas 'joy'
56.	/f/	The consonant /f/	**f**ãbrikas 'factory'
57.	/f"/	The soft consonant /f/	**f**igūrà 'figure', 'shape'
58.	/x/	The consonant /x/	**ch**òras 'chorus', 'choir'
59.	/x"/	The soft consonant /x/	**ch**èmija 'chemistry'
60.	/h/	The consonant /h/	**h**armònija 'harmony'
61.	/h"/	The soft consonant /h/	**h**iacìntas 'hyacinth'

62.	/s/	The consonant /s/	**s**áulė 'sun'
63.	/s"/	The soft consonant /s/	vaĩ**s**ius 'fruit', 'fetus'
64.	/š/	The consonant /š/	**š**akà 'branch'
65.	/š"/	The soft consonant /š/	**š**iáudas 'straw'
66.	/z/	The consonant /z/	**z**ýlė 'tit'
67.	/z"/	The soft consonant /z/	**z**irzė́ti 'to whine'
68.	/ž/	The consonant /ž/	**ž**vãkė 'candle'
69.	/ž"/	The soft consonant /ž/	**ž**iógas 'grasshopper' **ž**ẽmė 'earth', 'soil'
70.	/j"/	The soft consonant /j/ (in Lithuanian, 'j' is always soft)	**j**ū́ra 'sea' **j**ìs 'he'
71.	/l/	The consonant /l/	vá**l**sas 'waltz'
72.	/l˜/	The stressed consonant /l/	vi**l̃**kas 'wolf'
73.	/l"/	The soft consonant /l/	va**l**ià 'will'
74.	/l"˜/	The soft stressed consonant /l/	gu**l̃**ti 'to go to bed', 'to lie down'
75.	/m/	The consonant /m/	ã**m**atas 'handicraft'
76.	/m˜/	The stressed consonant /m/	li**m̃**palas 'adhesive '
77.	/m"/	The soft consonant /m/	s**m**ẽgenys 'brain'
78.	/m"˜/	The soft stressed consonant /m/	ka**m̃**štis 'cork'
79.	/n/	The consonant /n/	**n**ãmas 'house'
80.	/n˜/	The stressed consonant /n/	i**ñ**karas 'anchor'
81.	/n"/	The soft consonant /n/	**n**ėšti 'to carry along'
82.	/n"˜/	The soft stressed consonant /n/	le**ñ**ktis 'to bend', ' to bow'
83.	/r/	The consonant /r/	**r**ãtas 'wheel' be**r**žas 'birch'
84.	/r˜/	The stressed consonant /r/	ga**r̃**sas 'sound'
85.	/r"/	The soft consonant /r/	k**r**iáušė 'pear'
86.	/r"˜/	The soft stressed consonant /r/	ki**r̃**tis 'stress', 'blow'
87.	/v/	The consonant /v/	**v**óras 'spider'
88.	/v"/	The soft consonant /v/	**v**iáuksėti 'to yelp'
89.	/j/	The consonant sound /j/ (appearing when pronouncing the diphthongs 'ai', 'ei' and 'ui')	á**i**das 'echo' mé**i**lė 'love'
90.	/j˜/	The stressed consonant sound /j/ (appearing when pronouncing the diphthongs 'ai, 'ei' and 'ui')	e**ĩ**smas 'traffic' zu**ĩ**kis 'rabbit'
91.	/w/	The vowel sound /u/ (appearing when pronouncing the diphthong 'au')	ká**u**las 'bone'
92.	/w˜/	The stressed vowel sound /u/ (appearing when pronouncing the diphthong 'au')	la**ũ**kas 'field'

Appendix B. The vowel phoneme signals

The plots of the vowel phoneme signals of duration 0.02 *s* are shown in Fig. 1 – Fig. 6. The phonemes were obtained from female utterances. These plots reveal the periodic character of the vowel phoneme signals.

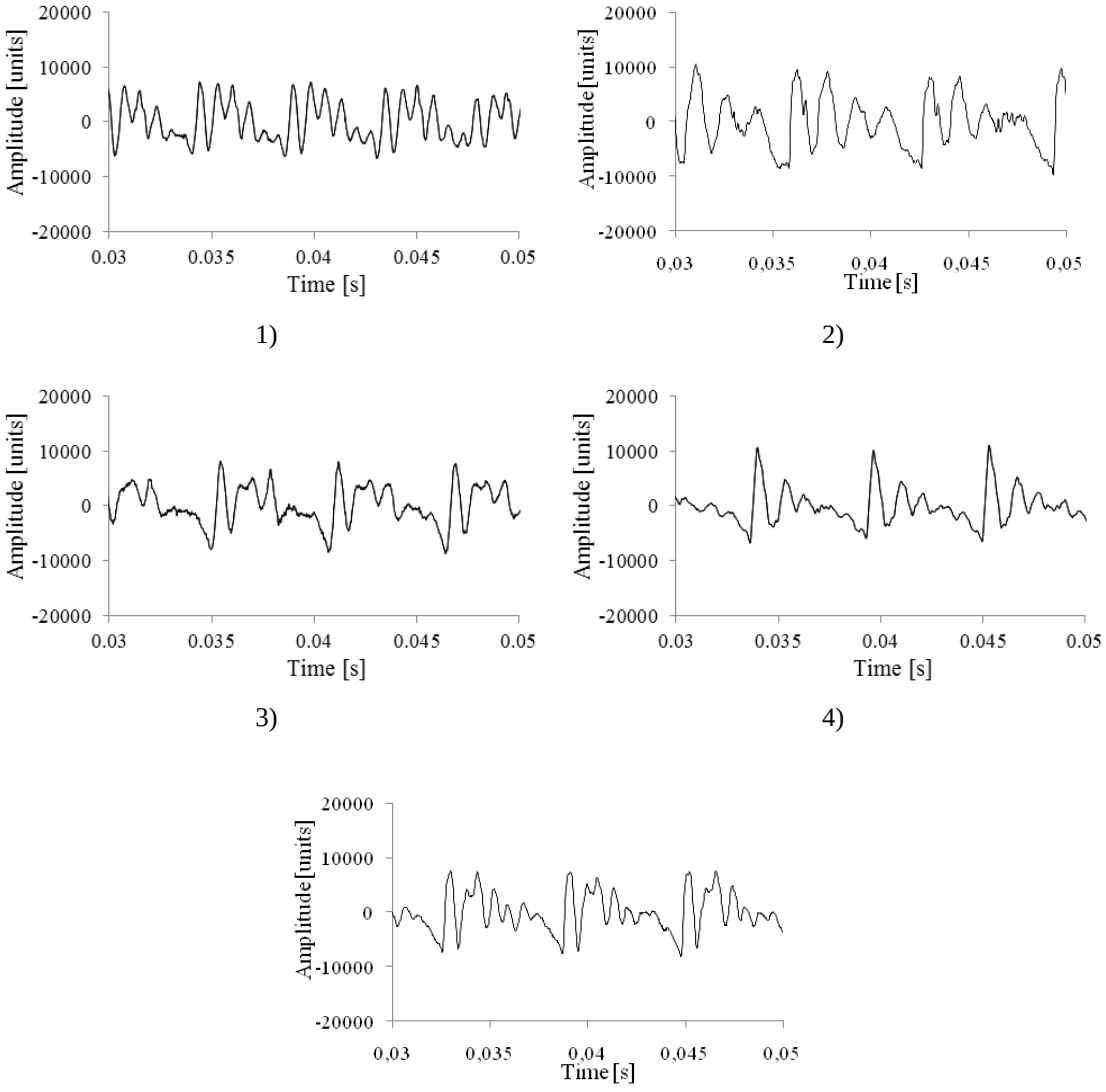


Fig. 1 *The plots of the vowel "a" phonemes: 1) the short unstressed vowel /a/ as in the word m**a**mà (mother), 2) the short stressed vowel /a`/ as in the word lazd**à** (stick), 3) the long unstressed vowel /a:/ as in the word dr**ą**sà (courage), 4) the long vowel stressed with the falling accent /a:ˊ/ as in the word k**á**rdas (sword), 5) the long vowel stressed with the rising accent /a:˜/ as in the word **ã**čiū (thank you)*

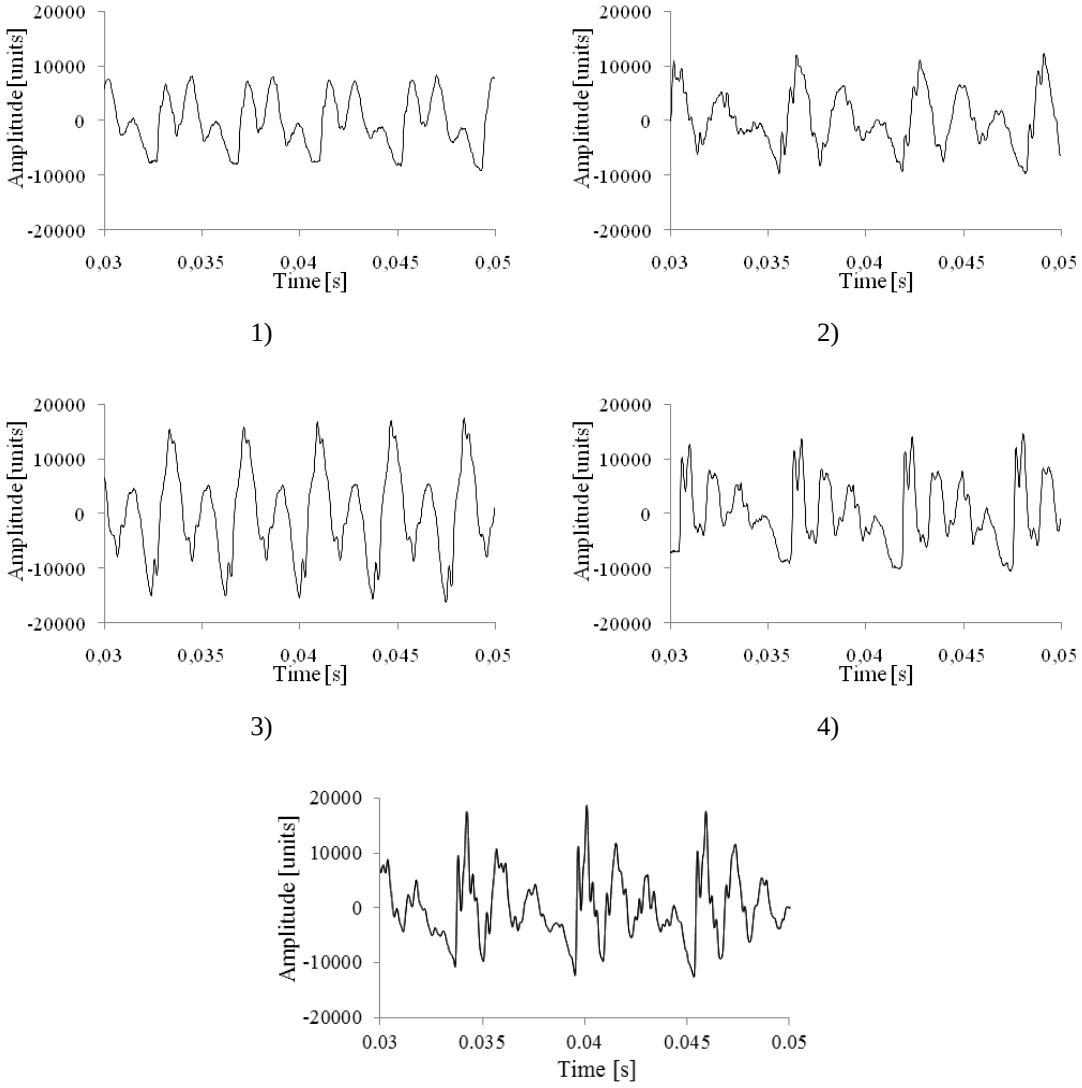


Fig. 2 *The plots of the vowel "e" phonemes: 1) the short unstressed vowel /e/ as in the word m**e**dãlis (medal), 2) the short stressed vowel /e`/as in the word sug**è**sti (turn bad), 3) the long unstressed vowel /e:/ as in the word gr**ę**žinỹs (well), 4) the long vowel stressed with the falling accent /e:´/ as in the word **é**rkė (mite), 5) the long vowel stressed with the rising accent /e:˜/ as in the word gyv**ẽ**nimas (life)*

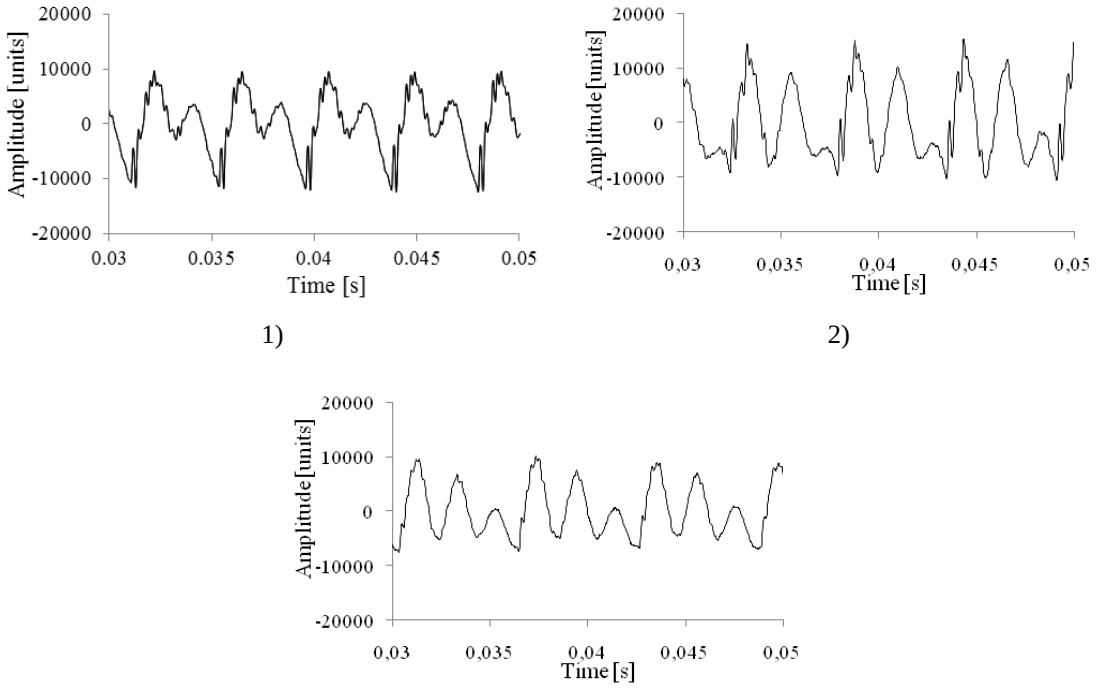
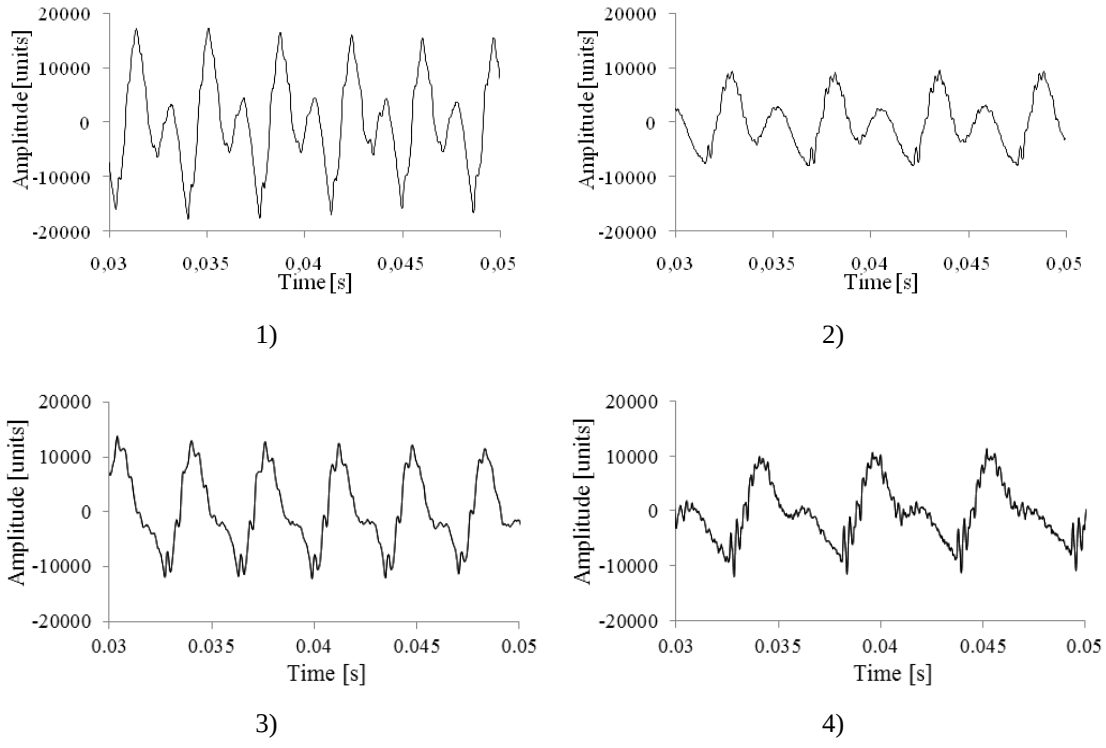
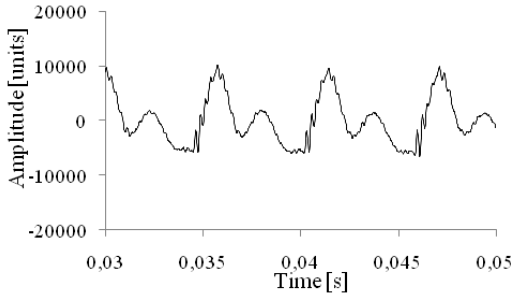


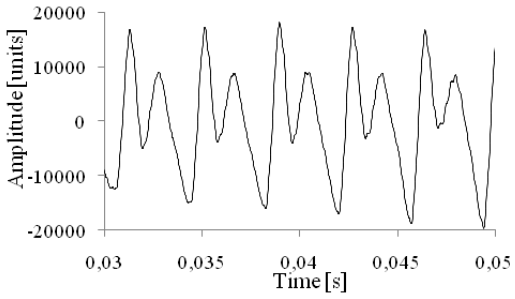
Fig. 3 *The plots of the vowel "ė" phonemes: 1) the long unstressed vowel /ė:/ as in the word kėdẽ (chair), 2) the long vowel stressed with the falling accent /ė:ˊ/ as in the word upėtakis (trout), 3) the long vowel stressed with the rising accent /ė:˜/ as in the word gėlẽ (flower)*



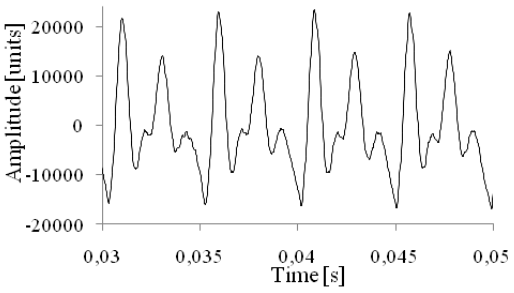


5)

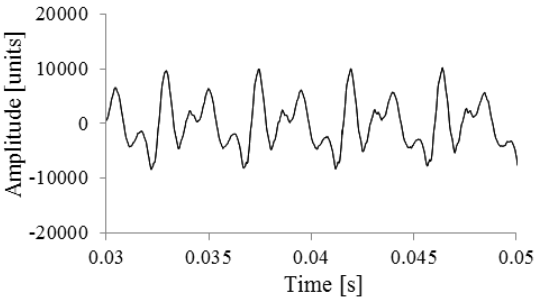
Fig. 4 *The plots of the vowel "i" phonemes: 1) the short unstressed vowel /i/ as in the word l**i**gà (disease), 2) the short stressed vowel /i`/ as in the word k**ì**škis (rabbit), 3) the long unstressed vowel /i:/ as in the word t**y**là (silence), 4) the long vowel stressed with the falling accent /i:ˊ/ as in the word r**ý**tas (morning), 5) the long vowel stressed with the rising accent /i:˜/ as in the word arkl**ỹ**s (horse)*



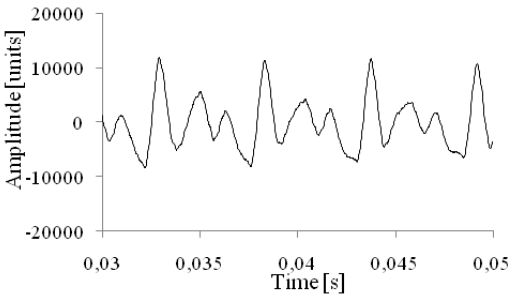
1)



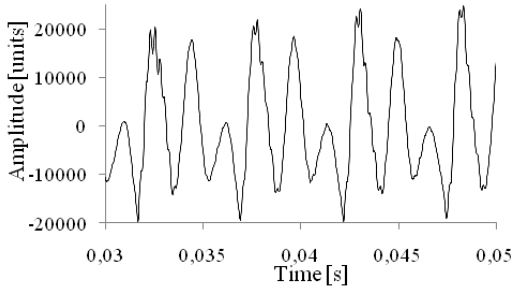
2)



3)

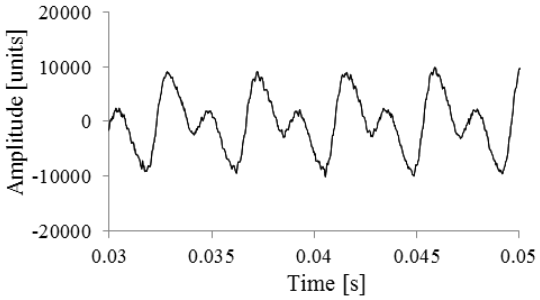


4)

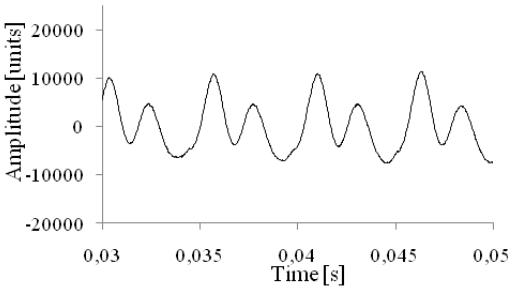


5)

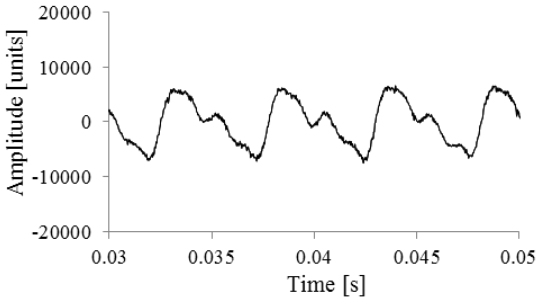
Fig. 5 *The plots of the vowel "o" phonemes: 1) the short unstressed vowel /o/ as in the word* ***o****žkà (she-goat), 2) the short stressed vowel /o`/ as in the word ch****ò****ras (choir), 3) the long unstressed vowel /o:/ as in the word k****o****vótojas (fighter), 4) the long vowel stressed with the falling accent /o:´/ as in the š****ó****nas (side), 5) the long vowel stressed with the rising accent /o:˜/ as in the word Ad****õ****mas (Adam)*



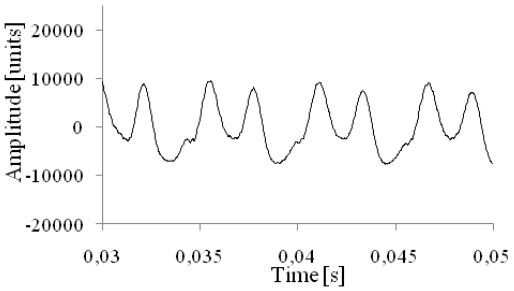
1)



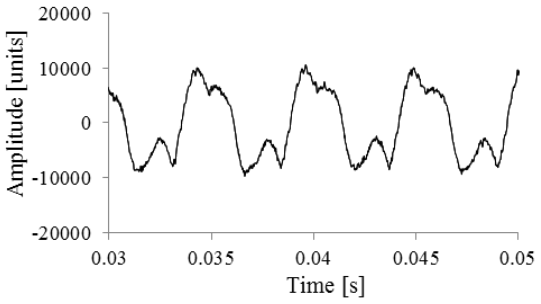
2)



3)



4)

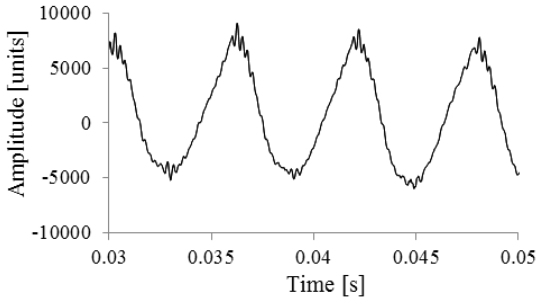


5)

Fig. 6 *The plots of the vowel "u" phonemes: 1) the short unstressed vowel /u/ as in the word k**u**ltūrà (culture), 2) the short stressed vowel /u`/ as in the word **ù**pė (river), 3) the long unstressed vowel /u:/ as in the word k**ū**rinỹs (work), 4) the long vowel stressed with the falling accent /u:´/ as in the word l**ū́**pa (lip), 5) the long vowel stressed with the rising accent /u:˜/ as in the word m**ū̃**šis (battle)*

Appendix C. The semivowel phoneme signals

The plots of the semivowel phoneme signals of duration 0.02 *s* are shown in Fig. 7 – Fig. 12. The phonemes were obtained from female utterances. These plots reveal the periodic character of the semivowel phoneme signals.



5)

Fig. 7 *The plot of the soft semivowel /j"/ phoneme as in the word j**ū́**ra (sea)*

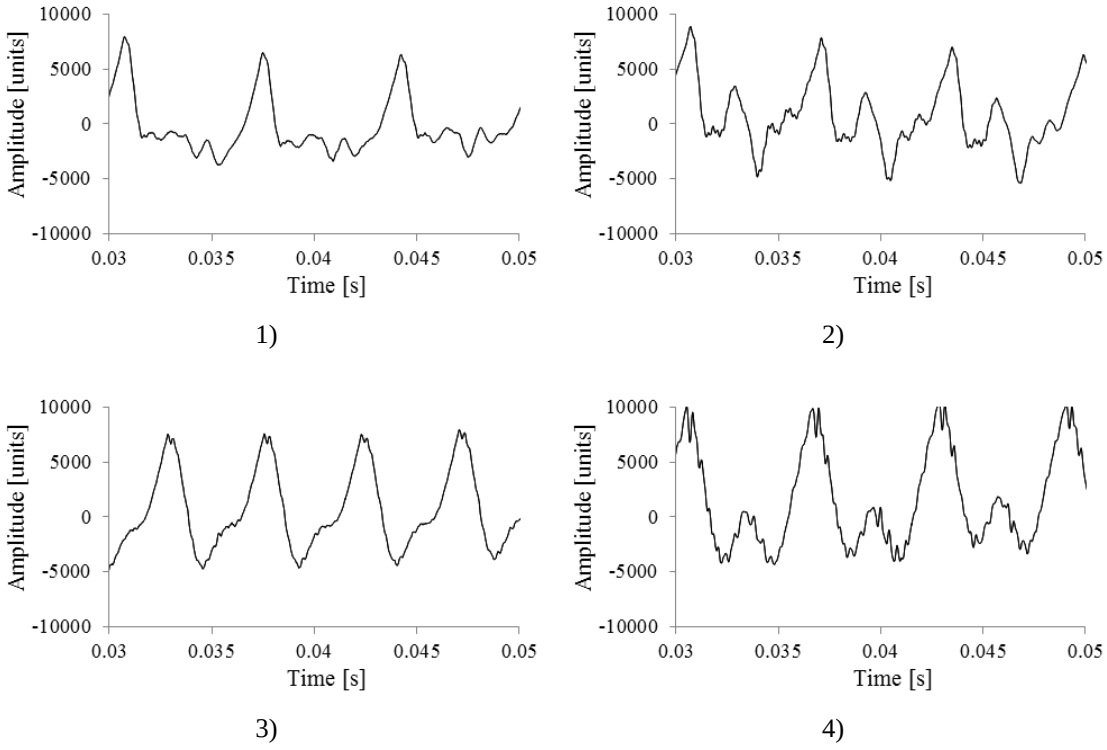
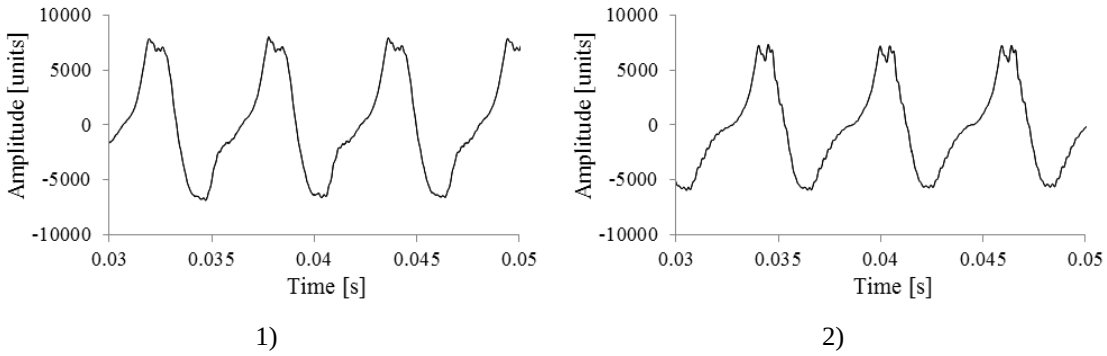
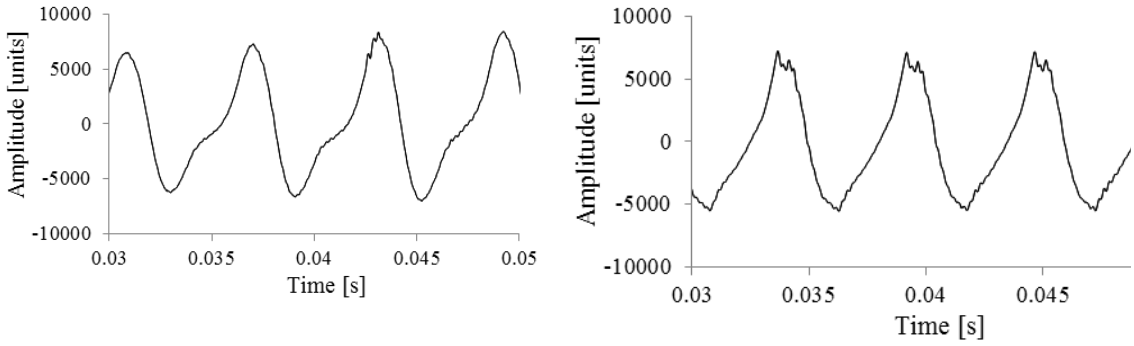


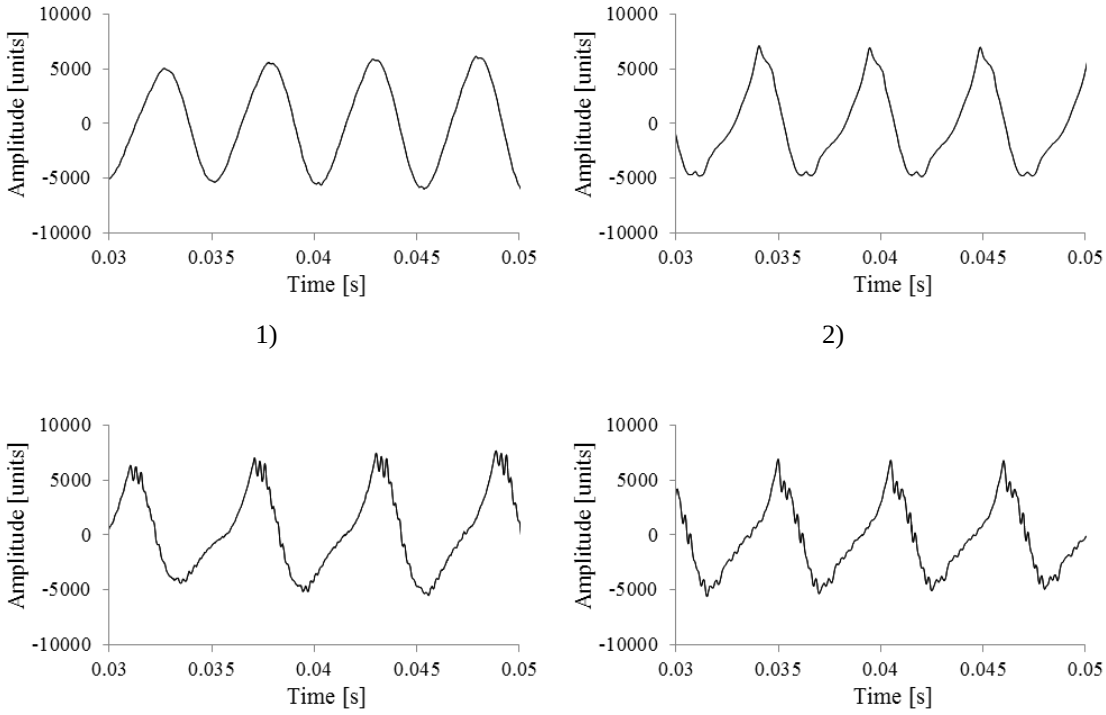
Fig. 8 *The plots of the semivowel "l" phonemes: 1) the unstressed semivowel /l/ as in the word válsas (waltz), 2) the stressed semivowel /l˜/ as in the word viĺkas (wolf), 3) the soft unstressed semivowel /l"/ as in the word valià (will), 4) the soft stressed semivowel /l"˜/ as in the word guĺti (to go to bed)*





3) 4)

Fig. 9 *The plots of the semivowel "m" phonemes: 1) the unstressed semivowel /m/ as in the word ã**m**atas (handicraft), 2) the stressed semivowel /m˜/ as in the word li**m̃**palas (adhesive), 3) the soft unstressed semivowel /m"/ as in the word s**m**ẽgenys (brain), 4) the soft stressed semivowel /m"˜/ as in the word ka**m̃**štis (cork)*



1) 2)

3) 4)

Fig. 10 *The plots of the semivowel "n" phonemes: 1) the unstressed semivowel /n/ as in the word* **n**ãmas (house), *2) the stressed semivowel /n˜/ as in the word* i**ñ**karas (anchor), *3) the soft unstressed semivowel /n"/ as in the word **n**ėšti (to carry along), 4) the soft stressed semivowel /n"˜/ as in the word le**ñ**ktis (to bend)*

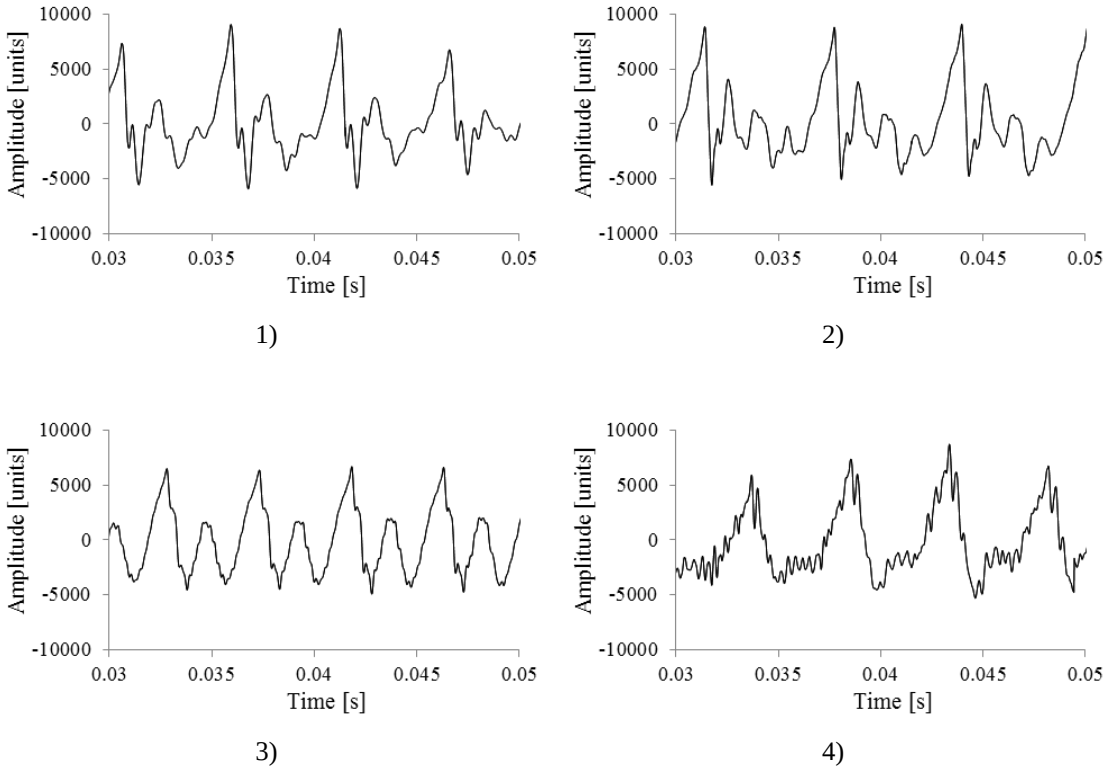


Fig. 11 *The plots of the semivowel "r" phonemes: 1) the unstressed semivowel /r/ as in the word* ***r****ãtas (wheel), 2) the stressed semivowel /r˜/ as in the word ga****r****˜sas (sound), 3) the soft unstressed semivowel /r"/ as in the word k****r****iáušė (pear), 4) the soft stressed semivowel /r"˜/ as in the word ki****r****˜tis (stress, blow)*

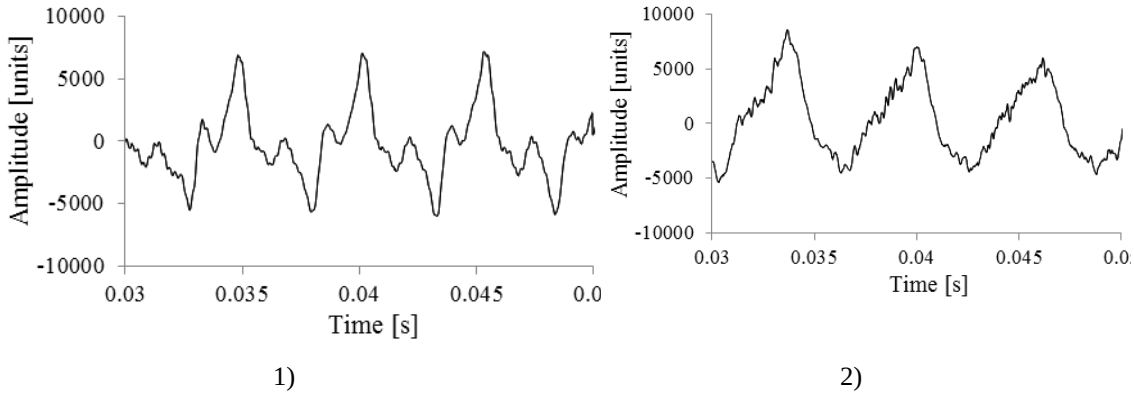


Fig. 12 *The plots of the semivowel "v" phonemes: 1) the unstressed semivowel /v/ as in the word* ***v****óras (spider), 2) the soft unstressed semivowel /v"/ as in the word* ***v****iáuksėti (to yelp)*

Gražina Pyž

ANALYSIS AND SYNTHESIS OF LITHUANIAN PHONEME DYNAMIC SOUND MODELS

Doctoral Dissertation

Technological Sciences,
Informatics Engineering (07T)

Gražina Pyž

LIETUVIŠKŲ FONEMŲ DINAMINIŲ MODELIŲ ANALIZĖ IR SINTEZĖ

Daktaro disertacija

Technologijos mokslai,
Informatikos inžinerija (07T)