INFORMATICA, 2011, Vol. 22, No. 3, 411–434



# Modelling of Lithuanian Speech Diphthongs

Gražina PYŽ[1], Virginija ŠIMONYTĖ[2], Vytautas SLIVINSKAS[2]
[1]*Vilnius University Institute of Mathematics and Informatics*
*Akademijos 4, LT-08663 Vilnius, Lithuania*
[2]*Vilnius Pedagogical University, Faculty of Mathematics and Informatics*
*Studentų 39, LT-08106 Vilnius, Lithuania*
*e-mail: grazinute123@gmail.com, virginija.simonyte@vpu.lt, vytautas.slivinskas@vpu.lt*

Received: January 2011; accepted: May 2011

**Abstract.** The goal of the paper is to get a method of Lithuanian speech diphthong modelling. We use a formant-based synthesizer for this modelling. The second order quasipolynomial has been chosen as the formant model in time domain. A general diphthong model is a multi-input and single-output (MISO) system, that consists of two parts where the first part corresponds to the first vowel of the diphthong and the second one – to the other vowel. The system is excited by semi-periodic impulses with a smooth transition from one vowel to the other. We derived the parametric input-output equations in the case of quasipolynomial formants, defined a new notion of the convoluted basic signal matrix, derived parametric minimization functional formulas for the convoluted output data. The new formant parameter estimation algorithm for convoluted data, based on Levenberg–Marquardt approach, has been derived and its stepwise form presented. Lithuanian diphthong /ai/ was selected as an example. This diphthong was recorded with the following parameters: PCM 48 kHz, 16 bit, stereo. Two characteristic pitches of the vowels /a/ and /i/ have been chosen. Equidistant samples of these pitches have been used for estimating parameters of MISO formant models of the vowels. Transition from the vowel /a/ to the vowel /i/ was achieved by changing excitation impulse amplitudes by the arctangent law. The method was 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 diphthongs. The magnitude and phase responses only have shown small differences.

**Keywords:** Lithuanian diphthongs, modelling, MISO system, Levenberg–Marquardt approach, formant, quasipolynomial model, parameter estimation, speech synthesis.

## 1. Introduction

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. Research groups work in Vilnius at the Institute of Mathematics and Informatics and 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 analyzed 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 of a framework for choosing a set of syllables and phonemes for Lithuanian speech recognition (Laurinčiukaitė and Lipeika, 2007), using of the formant features in the dynamic time warping based recognition of isolated Words (Lipeika and Lipeikienė, 2008; Lipeika, 2010). Experts at Vilnius University Faculty of Mathematics and Informatics consider text-to-speech synthesis of Lithuanian language (Kasparaitis, 2001), Lithuanian speech recognition using the English recognizer (Kasparaitis, 2008), text independent speaker identification (Bastys *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). 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), 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). Other fields (some of them are closely related with speech recognition) of Lithuanian language and speech engineering are noisy speech intelligibility enhancement (Kazlauskas, 1999), transcribing of the Lithuanian text (Kasparaitis, 1999; Skripkauskas and Telksnys, 2006), automatic stressing of the Lithuanian text (Kasparaitis, 2000), coding and transmission of voice signals (Kajackas and Anskaitis, 2009), the Lithuanian language machine translation (Šveikauskienė, 2005).

Lithuanian speech synthesis is one of the tasks of Lithuanian speech digital processing. In order to solve the problem of Lithuanian speech synthesis, it is necessary to develop mathematical models for Lithuanian speech sounds. Developing of the diphthong models is a part of this problem.

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). There exist two types of diphthongs in Lithuanian language that are made of two vowels: compound diphthongs (*ai, au, ei, ui, eu, oi, ou*) and complex diphthongs (*ie, uo*). A complex speech sound that begins with a short vowel (*i, e, u, a*) and ends with a sonant (*l, r, m, n*) within the same syllable is called a mixed diphthong (Garšva, 2001). The number of vowels and compound/complex diphthongs or mixed 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.

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

Diphthong modelling is important for solving a text-to-speech (TTS) problem. TTS problem arises in various applications such as services for the hearing impaired, reading email aloud, etc. A TTS system is defined as a system that takes a sequence of words as input and converts it into speech (SIL, 2004). 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 analyze and model durational patterns of natural speech in order to achieve an improved naturalness of synthetic speech".

A method for formant parameter extraction from a labelled single speaker database for use in a synthesis system is examined in Mannell (1998). This paper considers speech synthesis by the concatenation of formant parameter diphones. A word "diphone" can be derived from two Greek words "di" that means "two" and "phonos" that means "sound". A diphone is defined as an adjacent pair of phones.

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 distinguish 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 unanalyzed 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.

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

Two main speech synthesis types – concatenation (Hopcroft and Ullman, 1979; Kasparaitis, 2001) and formant (Holmes and Holmes, 2001; Slivinskas and Šimonytė, 2007) – are discussed in the literature. The formant synthesis type is based on the formant model. The main formant model problem is a strong synthetic shade of the synthesized signal (Ringys and Slivinskas, 2009).

The objective of this paper is to develop a method of Lithuanian speech diphthong modelling. We use the formant synthesizer method. Each formant is responsible for a certain band of the frequency response. In general case, the formants of the diphthong vowels are different. The first problem is as follows: what functions must be used to describe formants in time domain. A formant can be selected as a damped sinusoid. Such an

approach is often used in various language formant synthesizers. The sounds produced by such synthesizers, however, have synthetic character. Ringys and Slivinskas (2010) used a quasipolynomial function to describe formants. They obtained sufficiently naturally sounding vowel sounds. Following them, we choose a quasipolynomial function as a mathematical tool for formant description.

The second problem is to estimate formant parameters in the case of a diphthong. Ringys and Slivinskas (2010) used the Levenberg–Marquardt approach and an algorithm derived in Šimonytė and Slivinskas (1997) for vowel formant parameter estimation. They, however, did not take into account that the data is convoluted. Using the above mentioned algorithm significant errors are obtained. These errors are not essential in vowel modelling, they, however, can distort the overall diphthong model. Therefore we need to develop a quasypolinomial formant parameter estimation algorithm that can be used for convoluted data. The data can be deconvolved using the assumption that the formant decays in three periods.

The third problem is to choose the model type of a system producing the diphthong. If the system is SISO (Single-Input and Single-Output), then all the model formants are excited in the same way. As a result, we get the same output periods. Modelling of a natural language requires the use of MISO (Multiple-Input and Single-Output) system (Ringys and Slivinskas, 2010). In the paper, we use a MISO system consisting of two parts. The first one is intended for the first vowel modelling, and the other – for the second vowel modelling. All the formants influence the transition area.

Looking at the recorded diphthong vowel plots, one can notice that the periods (even adjacent) are not completely identical. To obtain a model for such a vowel signal, it is necessary to use a MISO system where each channel (SISO system) is responsible for one frequency band and is related to a single formant. If the formants are excited not in the same way, we can get more natural synthesized diphthong sounding.

The fourth problem is selection of the MISO system input. To solve this problem, we need to filter the diphthong signal in the formant intervals and calculate the output signal maximums. In this way, we obtain the input amplitude dynamics.

The fifth problem is selection of the area of the transition from the first vowel to the second vowel. We have to choose a mathematical function describing this transition.

This paper is the first one in the field of Lithuanian diphthong modelling. Our aim is not to model all the Lithuanian diphthongs but to propose a method for their modelling. Therefore we present only a single example – modelling of the diphthong /ai/.

The paper is structured as follows: Section 2 deals with derivation of the parametric input-output equations in the case of quasipolynomial formants. We define a new notion of the convoluted basic signal matrix and derive parametric minimization functional formulas for the convoluted output data. In Section 3, we develop formant parameter estimation algorithm for convoluted data. We present a stepwise form of this algorithm in Section 4. In Section 5, we give not only an example, but also the methodological outline for diphthong modelling.

## 2. A MISO System as a Speech Diphthong Model

Consider a MISO (multiple-input and single-output) discrete time system that is made of $K$ causal SISO (single-input and single-output) systems connected in parallel (Huang *et al.*, 2010; Ringys and Slivinskas, 2010). We assume that $K = K_1 + K_2$ where $K_1$ equals the number of SISO systems corresponding to the first vowel of the diphthong, $K_2$ – to the number of those SISO systems corresponding to the second vowel.

Let

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

denote the impulse responses, and

$$\begin{aligned}\mathbf{u}_1 &= \big(\dots, u_1(-1), u_1(0), u_1(1), \dots\big),\\ &\vdots\\ \mathbf{u}_K &= \big(\dots, u_K(-1), u_K(0), u_K(1), \dots\big)\end{aligned} \tag{2}$$

– the inputs of the corresponding SISO systems. Denote by

$$\mathbf{y}_k = \big(\dots, y_k(-1), y_k(0), y_k(1), \dots\big), \quad k = 1, \dots, K \tag{3}$$

the output of the $k$th SISO system.

The impulse response $\mathbf{H}$ of the MISO system is the following vector of sequences:

$$\mathbf{H} = \begin{bmatrix}\mathbf{h}_1\\ \vdots\\ \mathbf{h}_K\end{bmatrix}, \tag{4}$$

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

We consider such a MISO system whose output $\mathbf{y}$ is equal to the sum of the inputs of the SISO systems:

$$\mathbf{y} = \mathbf{y}_1 + \cdots + \mathbf{y}_K. \tag{5}$$

Since

$$\mathbf{y}_k = u_k * h_k, \tag{6}$$

where $*$ denotes the convolution operation

$$\mathbf{u}_k * \mathbf{h}_k = \sum_{i=0}^{\infty} u_k(n-i)h_k(i), \tag{7}$$

we have

$$\mathbf{y} = u_1 * h_1 + \ldots + u_K * h_K. \tag{8}$$

A diagram of such a MISO system is shown in Fig. 1.

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

$$\boldsymbol{\delta}(n) = \begin{cases} [1, 1, \ldots, 1]^T, & n = 0, \\ [\underbrace{0, 0, \ldots, 0}_{K}]^T, & n \neq 0. \end{cases} \tag{9}$$

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

$$\mathbf{y}(n) = \sum_{k=1}^{K} \mathbf{h}_k(n), \quad n \geqslant 0. \tag{10}$$

We assume that all the $\mathbf{h}_k$ can be represented by second order quasipolynomials (they are defined below). Quasipolynomials are slightly more sophisticated than damped sinusoids.

A formant is defined as a vocal tract resonance (Rosner and Pickering, 1994; Fry, 1977). The term "formant" is often used to refer to a peak in a frequency-domain representation (Rosner and Pickering, 1994). In time-domain representation, a formant can be described as a damped sinusoid (Fant, 1970; Cook, 2002).

This, however, is true as a special case only. It has been shown that a formant has a more general form Slivinskas and Šimonytė (1990). In this book, the rigorous definitions of a signal, its order, formant of a signal were presented, and it was proved that in general case the formant of a signal is a quasipolynomial. In this paper, we restrict ourselves to second order quasipolynomials from practical considerations. Such a quasipolynomial is a damped exponential multiplied by a second order polynomial whose coefficients vary according to a sinusoidal law with the same frequency and different amplitudes and

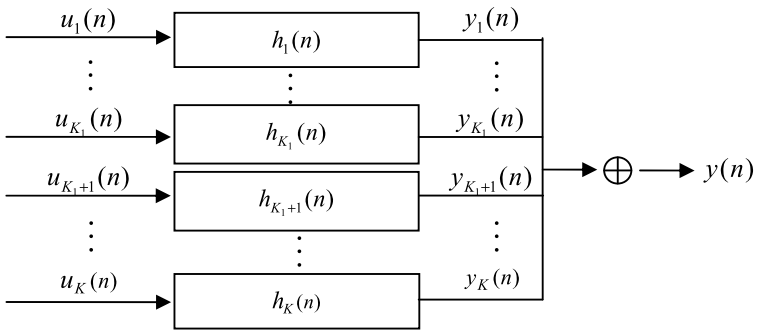


Fig. 1. A MISO system diagram.

phases. The mathematical description of a second order continuous-time quasipolynomial is given below:

$$\begin{aligned} s(t) = \ & \mathrm{e}^{\lambda t}\big(a_1 \sin(2\pi f t + \varphi_1) + a_2 t \sin(2\pi f t + \varphi_2) \\ & + a_3 t^2 \sin(2\pi f t + \varphi_3)\big), \end{aligned} \tag{11}$$

where $t$ is a nonnegative real number $t \in R^+ \cup \{0\},\ \lambda < 0$ – the damping factor, $f$ – the frequency, $a_k$ – amplitude, $\varphi_k (-\pi \leqslant \varphi_k < \pi)$ – 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 formant:

$$h_k(n) = s(n\Delta t), \quad n = 0, 1, 2, \ldots, \tag{12}$$

where the formant $s(t)$ is defined by (11) 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 (11) and (12) that

$$\begin{aligned} & \mathrm{h}_k(0) = a_{k1} \sin(\varphi_{k1}), \\ & \mathrm{h}_k(\Delta t) = \mathrm{e}^{\Delta t \lambda_k}\big(a_{k1} \sin(2\pi f_k \Delta t + \varphi_{k1}) + a_{k2} \Delta t \sin(2\pi f_k \Delta t + \varphi_{k2}) \\ & \qquad + a_{k3} \Delta t^2 \sin(2\pi f_k \Delta t + \varphi_{k3})\big), \\ & \mathrm{h}_k(2\Delta t) = \mathrm{e}^{2\Delta t \lambda_k}\big(a_{k1} \sin(2\pi f_k 2\Delta t + \varphi_{k1}) + a_{k2} 2\Delta t \sin(2\pi f_k 2\Delta t + \varphi_{k2}) \\ & \qquad + a_{k3} 2^2 \Delta t^2 \sin(2\pi f_k 2\Delta t + \varphi_{k3})\big), \\ & \vdots \\ & \mathrm{h}_k((N-1)\Delta t) = \mathrm{e}^{(N-1)\Delta t \lambda_k}\big(a_{k1} \sin(2\pi f_k (N-1)\Delta t + \varphi_{k1}) \\ & \qquad + a_{k2}(N-1)\Delta t \sin(2\pi f_k (N-1)\Delta t + \varphi_{k2}) \\ & \qquad + a_{k3}(N-1)^2 \Delta t^2 \sin(2\pi f_k (N-1)\Delta t + \varphi_{k3})\big). \end{aligned} \tag{13}$$

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

$$\mathbf{h}_k^N = \big[h_k(0), h_k(\Delta t), h_k(2\Delta t), \ldots, h_k((N-1)\Delta t)\big]^T. \tag{14}$$

Let

$$\begin{aligned} & \Omega_k = 2\pi f_k \Delta t, \\ & \Lambda_k = \lambda_k \Delta t, \\ & A_{k1} = a_{k1}, \\ & A_{k2} = a_{k2} \Delta t, \\ & A_{k3} = a_{k3} \Delta t^2, \quad k = 1, \ldots, K. \end{aligned} \tag{15}$$

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

$$\mathbf{y}^N = \big[y(0), y(1), y(2), \ldots, y(N-1)\big]^T. \tag{16}$$

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

$$\boldsymbol{\Psi}_k = [\boldsymbol{\Psi}_{k1}\ \boldsymbol{\Psi}_{k2}\ \boldsymbol{\Psi}_{k3}], \tag{17}$$

where

$$\boldsymbol{\Psi}_{ki} = \begin{bmatrix} \delta(i-1) & 0 \\ c_{ki} & s_{ki} \end{bmatrix} \qquad \left(\delta(i) = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0 \end{cases}\right), \tag{18}$$

with

$$\begin{aligned} c_{ki} = \big[&\mathrm{e}^{\Lambda_k} \cos\Omega_k, 2^{i-1}\mathrm{e}^{2\Lambda_k} \cos 2\Omega_k, \ldots, \\ &(N-1)^{i-1}\mathrm{e}^{(N-1)\Lambda_k} \cos(N-1)\Omega_k\big]^T, \end{aligned} \tag{19}$$

$$\begin{aligned} s_{ki} = \big[&\mathrm{e}^{\Lambda_k} \sin\Omega_k, 2^{i-1}\mathrm{e}^{2\Lambda_k} \sin 2\Omega_k, \ldots, \\ &(N-1)^{i-1}\mathrm{e}^{(N-1)\Lambda_k} \sin(N-1)\Omega_k\big]^T, \end{aligned} \tag{20}$$

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

$$\begin{aligned} \boldsymbol{\alpha}_k = \big[&A_{k1}\sin(\varphi_{k1}), A_{k1}\cos(\varphi_{k1}), A_{k2}\sin(\varphi_{k2}), A_{k2}\cos(\varphi_{k2}), \\ &A_{k3}\sin(\varphi_{k3}), A_{k3}\cos(\varphi_{k3})\big]^T. \end{aligned} \tag{21}$$

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

$$\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1 \mid \boldsymbol{\Psi}_2 \mid \ldots \mid \boldsymbol{\Psi}_K], \tag{22}$$

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

$$\boldsymbol{\alpha} = \big[\boldsymbol{\alpha}_1^T \mid \boldsymbol{\alpha}_2^T \mid \ldots \mid \boldsymbol{\alpha}_K^T\big]^T. \tag{23}$$

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

$$\mathbf{y}^N = \boldsymbol{\Psi}_1\boldsymbol{\alpha}_1 + \cdots + \boldsymbol{\Psi}_K\boldsymbol{\alpha}_K = \boldsymbol{\Psi} \cdot \boldsymbol{\alpha}. \tag{24}$$

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

$$\boldsymbol{\alpha} = \boldsymbol{\Psi}^+ \cdot \mathbf{y}^N, \tag{25}$$

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 $z$ denotes the common shift operator. Then from (4) we can write

$$z\mathbf{H} = \begin{bmatrix} z\mathbf{h}_1 \\ \vdots \\ z\mathbf{h}_K \end{bmatrix}, \tag{26}$$

where

$$z\mathbf{h}_k = \big(h_k(1), h_k(2), h_k(3), \dots\big), \quad k = 1, \dots, K \tag{27}$$

is the shifted sequence.

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

$$\boldsymbol{\delta}(n),\ \boldsymbol{\delta}(n-M),\ \boldsymbol{\delta}(n-2M),\ \dots, \tag{28}$$

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

$$z^k\mathbf{H} = \left.\begin{bmatrix} 0, & 0, & 0, & \dots \\ \vdots & & & \\ 0, & 0, & 0, & \dots \end{bmatrix}\right\} K \text{ rows}, \tag{29}$$

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

$$\begin{aligned}
&\mathrm{y}(n) = 0 \quad \text{for } n < 0, \\
&\mathrm{y}(n) = \sum_{k=1}^{K} \mathrm{h}_k(n) \quad \text{for } \leqslant n < M, \\
&\mathrm{y}(n) = \sum_{k=1}^{K} \big(\mathrm{h}_k(n) - \mathrm{h}_k(n-M)\big) \quad \text{for } M \leqslant n < 2M, \\
&\mathrm{y}(n) = \sum_{k=1}^{K} \big(\mathrm{h}_k(n) + \mathrm{h}_k(n-M) + \mathrm{h}_k(n-2M)\big) \\
&\quad \text{for } l \cdot M \leqslant n < (l+1) \cdot M, \quad l = 2, 3, \dots .
\end{aligned} \tag{30}$$

Let $\boldsymbol{\theta} = [\Lambda_1, \Omega_1, \dots, \Lambda_K, \Omega_K]$. Introduce the following $M \times 6K$ matrix $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\boldsymbol{\theta})$ (we call it the convoluted matrix):

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}(1 : M, :) + \boldsymbol{\Psi}(M + 1 : 2M, :) + \boldsymbol{\Psi}(2M + 1 : 3M, :), \tag{31}$$

where $\boldsymbol{\Psi}(m : n, :)$ denotes a submatrix of the matrix $\boldsymbol{\Psi}$ 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 $\boldsymbol{\Phi}$

depends on the damping factors $\Lambda_k$ and the angular frequencies $\Omega_k$(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 $\mathbf{y}^M$:

$$\mathbf{y}^M = \big[y(0), y(1), y(2), \ldots, y(M-1)\big]^T. \tag{32}$$

Using (30), it is not difficult to check that

$$\mathbf{y}^M = \boldsymbol{\Phi} \cdot \boldsymbol{\alpha}. \tag{33}$$

Usually we measure the output $y(t)$ with errors. Therefore we must incorporate an error component into (33). This error is modelled by an additive white Gaussian noise. Collect all the noise values into a vector $\mathbf{e}$:

$$\mathbf{e} = \big[e(0), e(1), e(2), \ldots, e(N-1)\big]^T. \tag{34}$$

Then the vector $\mathbf{y}$ can be written in the following form:

$$\mathbf{y}^M = \boldsymbol{\Phi} \cdot \boldsymbol{\alpha} + \mathbf{e}. \tag{35}$$

We have to minimize the following functional:

$$r(\mathbf{a}, \boldsymbol{\theta}) = \big\|\mathbf{y}^M - \boldsymbol{\Phi}(\boldsymbol{\theta})\boldsymbol{\alpha}\big\|^2, \tag{36}$$

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:

$$r_2(\boldsymbol{\theta}) = \big\|P^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathbf{x}\big\|^2, \tag{37}$$

where

$$P^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})} = \mathrm{I}_M - \boldsymbol{\Phi}\big(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\big)^{-1}\boldsymbol{\Phi}^{\mathrm{T}} = \mathrm{I}_M - \boldsymbol{\Phi}\boldsymbol{\Phi}^{+}. \tag{38}$$

$P^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}$ 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 $\hat{\theta}$ of the parameter $\boldsymbol{\theta}$ that minimizes $r_2(\boldsymbol{\theta})$. Then the value $(\hat{\mathbf{a}},\hat{\boldsymbol{\theta}})$ where

$$\hat{\boldsymbol{a}} = \boldsymbol{\Phi}^{+}(\hat{\boldsymbol{\theta}})\mathbf{y} \tag{39}$$

minimizes the functional $r(\mathbf{a},\boldsymbol{\theta})$. We use Levenberg–Marquardt approach to minimize the functional (37). In the next section, we develop a new algorithm applied to the convoluted data. The algorithm for nonconvoluted data was developed in Šimonytė and Slivinskas (1997).

## 3. Parameter Estimation of the Diphthong Formant Synthesis Model

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

$$\boldsymbol{\theta}^{l+1} = \boldsymbol{\theta}^{l} - \left(\mathbf{V}^{T}(\boldsymbol{\theta}^{l})\mathbf{V}(\boldsymbol{\theta}^{l}) + c_{l}\mathbf{I}_{2K}\right)^{-1}\mathbf{V}^{T}\left(\boldsymbol{\theta}^{l}\right)\mathbf{b}\left(\boldsymbol{\theta}^{l}\right), \quad l = 0, 1, \ldots, \tag{40}$$

where

$$\mathbf{V}(\boldsymbol{\theta}) = \mathcal{D}\left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\right)\mathbf{y} \tag{41}$$

is an $N \times 2K$ matrix (the symbol $\mathcal{D}$ stands for differentiation operation $\mathcal{D} = \frac{\partial}{\partial \boldsymbol{\theta}}, \boldsymbol{\theta}^{l}$ 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),

$$\mathbf{b}(\boldsymbol{\theta}) = \mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathbf{y} \tag{42}$$

is an $N \times 1$ 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 (40) 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 (40) computation algorithm in the explicit form using constructive matrix operations (addition, subtraction, multiplication, pseudoinverse, QR decomposition). Our data model is described in Section 2. One of such unconstructive operations in (40)–(41) is differentiation. In order to implement (41), we need to calculate $\mathcal{D}(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})})$ for our data model. This is a rather difficult task. Luckily there exists a formula that simplifies calculation of $\mathcal{D}(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})})$. It is shown in Golub and Pereyra (1973) that

$$\mathcal{D}\left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\right) = -\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathcal{D}(\boldsymbol{\Phi})\mathbf{B} - \left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}\mathcal{D}(\boldsymbol{\Phi})\mathbf{B}\right)^{T}, \tag{43}$$

where $\mathcal{D}(\boldsymbol{\Phi})$ is a three-dimensional tensor formed of $2KM \times 6K$ 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 $\boldsymbol{\Phi}$ with respect to the $k$th component $\theta_k$ of the parameter vector $\boldsymbol{\theta}$:

$$\mathbf{G}_k = \frac{\partial \boldsymbol{\Phi}}{\partial \theta_k}. \tag{44}$$

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

$$\left(\mathcal{D}(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})})\right)_k = -\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}G_k\mathbf{B} - \left(\mathbf{P}^{\perp}_{\boldsymbol{\Phi}(\boldsymbol{\theta})}G_k\mathbf{B}\right)^{T}. \tag{45}$$

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

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

$$\mathbf{B} = \mathbf{S}\big[T_1^{-1}\ \mathbf{0}_{6K\times(M-6K)}\big]\mathbf{Q}^{\mathrm{T}}. \tag{46}$$

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 $\mathbf{Q}$ (Golub and Pereyra, 1973):

$$P^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta})} = \mathbf{Q}^T \begin{bmatrix} \mathbf{0}_{6K\times 6K} & \mathbf{0}_{6K\times(M-6K)} \\ \mathbf{0}_{(M-6K)\times 6K} & \mathbf{I}_{M-6K} \end{bmatrix} \mathbf{Q}. \tag{47}$$

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 $\Lambda_k$ or angular frequency $\Omega_k$ 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 $k = 1, \ldots, K$:

$$\begin{aligned} \mathbf{G}_{2k-1} &= \frac{\partial \mathbf{\Phi}}{\partial \theta_{2k-1}} = \frac{\partial \mathbf{\Phi}}{\partial \Lambda_k}, \\ \mathbf{G}_{2k} &= \frac{\partial \mathbf{\Phi}}{\partial \theta_{2k}} = \frac{\partial \mathbf{\Phi}}{\partial \Omega_k}. \end{aligned} \tag{48}$$

It is not difficult to check that:

$$\begin{aligned} \mathbf{G}_{2k-1} &= \frac{\partial \mathbf{\Phi}}{\partial \Lambda_k} \\ &= \Bigg[ \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(k-1)},\ \mathbf{\Phi}(:, 6k-3),\ \mathbf{\Phi}(:, 6k-2),\ \mathbf{\Phi}(:, 6k-1), \\ &\qquad \mathbf{\Phi}(:, 6k),\ \left\{ \sum_{i=1}^{3} (im-1) \mathrm{e}^{(im-1)\Delta t \Lambda_k} \cos(im-1)\Omega_k \right\}_{m=1}^{M}, \\ &\qquad \left\{ \sum_{i=1}^{3} (im-1) \mathrm{e}^{(im-1)\Delta t \Lambda_k} \sin(im-1)\Omega_k \right\}_{m=1}^{M}, \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(K-k)} \Bigg], \end{aligned} \tag{49}$$

and

$$\mathbf{G}_{2k} = \frac{\partial \mathbf{\Phi}}{\partial \Omega_k}$$
$$= \Bigg[ \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(k-1)}, \mathbf{\Phi}(:, 6k-2), \mathbf{\Phi}(:, 6k-3), \mathbf{\Phi}(:, 6k),$$
$$\mathbf{\Phi}(:, 6k-1), -\left\{ \sum_{i=1}^{3} (im-1) \mathrm{e}^{(im-1)\Delta t \Lambda_k} \sin(im-1)\Omega_k \right\}_{m=1}^{M},$$
$$\left\{ \sum_{i=1}^{3} (im-1) \mathrm{e}^{(im-1)\Delta t \Lambda_k} \cos(im-1)\Omega_k \right\}_{m=1}^{M}, \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}_{6(K-k)} \Bigg]. \tag{50}$$

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

## 4. Formant Parameter Estimation Algorithm

All the formulas that we derived in Section 3 ((42)–(50)) are constructive. Using these formulas, we can present a stepwise algorithm for calculating formant parameter estimates from convoluted data.

Given:

1. The index $k$ of the formant under investigation.
2. The sequence **y** obtained from the pitch samples using the inverse Fourier transform in the frequency range corresponding to the $k$th formant.
3. The initial parameter vector $\boldsymbol{\theta}^0 = [\Omega_k^0 \ \Lambda_k^0]^T$.
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 percents $\varepsilon_{\min}$.
8. The initial signal estimation relative accuracy in percents $\varepsilon_{-1}$, 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 $l \geqslant l_{\max}$ 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 (17)–(20).
Step 2. Compute the convolution matrix $\mathbf{\Phi} = \mathbf{\Phi}_k(\boldsymbol{\theta}^l)$ by equation (31).
Step 3. Using the standard QR decomposition of the matrix $\mathbf{\Phi}$, find the general inverse matrix $\mathbf{B}$ according to formula (46).
Step 4. Compute the projector onto the noise subspace $P^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta}^l)}$ by formula (38).
Step 5. Find the projection of $\mathbf{y}$ onto the noise subspace by formula (42).
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 (37).
Step 7. If $(\varepsilon_l < \varepsilon_{l-1})$ then
$c_l = c_{l-1}/10$
Go to Step 8
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 $\Lambda_k$ and $\Omega_k$ using formulas (49)–(50).
Step 9. Find the derivative of the projector $D(\mathbf{P}^{\perp}_{\mathbf{\Phi}(\boldsymbol{\theta}^l)})$ according to equation (45).
Step 10. Calculate the matrix $V(\boldsymbol{\theta}^l)$by formula (41).
Step 11. Compute the parameter vector $\boldsymbol{\theta}^{l+1}$ by (40) 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 $\mathbf{\Psi}$.

## 5. Experiment Results

We considered an utterance of the Lithuanian word “laimė” (“happiness“). Its duration was 1.34 s. This utterance was recorded to a wav file of the audio format with the following parameters: PCM 48 kHz; 16 bit; stereo. Since both channel data sets were identical, we examined the data from a single channel. We considered not all of these samples, but a part of them corresponding to the compound diphthong /ai/. The duration of this part was 0.47 s. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ be equidistant samples of the diphthong /ai/ where $N = 0.47 \cdot 48000 = \mathbf{22\,560}$ samples. These samples are shown in Fig. 2. It is not difficult to see that this discrete signal exhibit 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.

We select two periods corresponding to the vowel /a/ and the vowel /i/. We choose the period in such a way that: (1) the first sample of the period be as near as possible to the

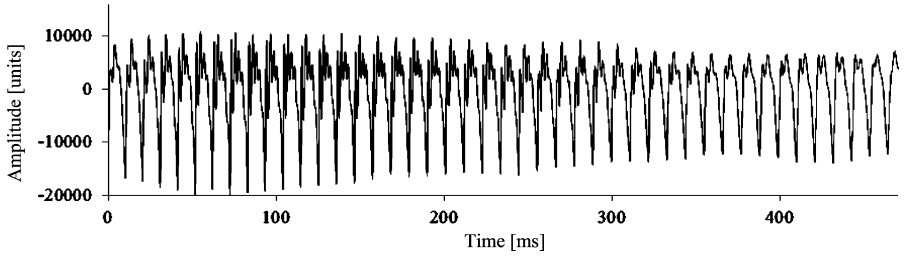


Fig. 2. The samples of a discretized version of the diphthong “ai” of the Lithuanian word “laimė”.

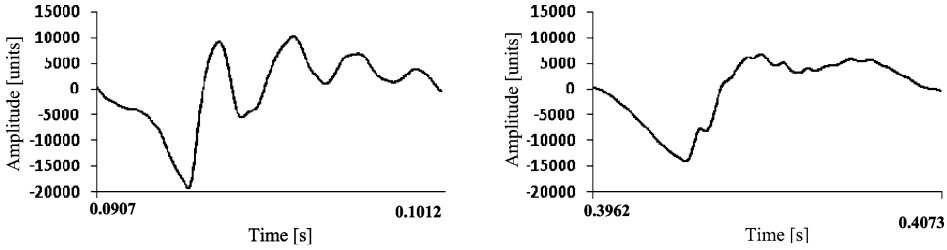


Fig. 3. The pitch corresponding (a) to the vowel “a” and (b) to the vowel “i”.

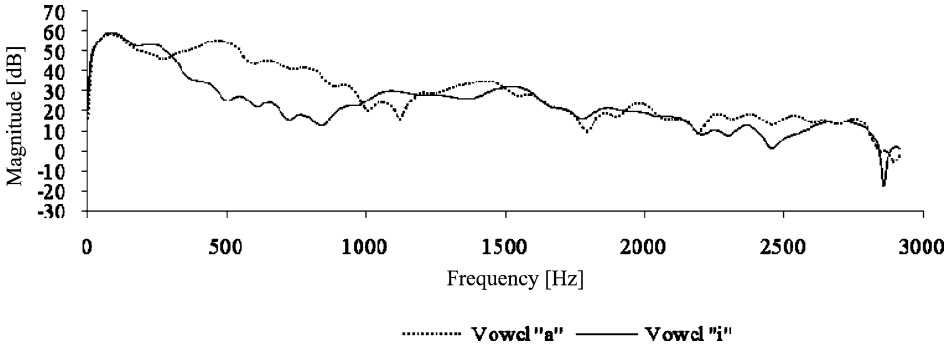


Fig. 4. The magnitude response of the vowels “a” and “i”.

zero value; (2) the energy of the first half of the period be larger than that of the second half. We call the selected periods pitches. The pitch corresponding to the vowel /a/ is the 10th period of the diphthong /ai/, and the pitch corresponding to the vowel /i/ is the 38th period. These pitches are shown in Fig. 3.

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. 4.

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

Table 1
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 | |

After analyzing 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 1.

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 (11). Estimation was done using Levenberg–Marquardt method described in Section 3 of this paper.

We will describe the procedure of the quasipolynomial model obtaining from the data of a single formant interval in more detail. Consider the 3rd formant interval of the vowel /a/. First, we select the initial estimates of the damping factor and angular frequency: $\Lambda_0 = -0.02,\ \Omega_0 = 0.046$ [rad/sample]. We use these estimates in the iterative procedure described by (40) where $\boldsymbol{\theta} = [\Lambda, \Omega]^T$. This procedure is repeated until the number

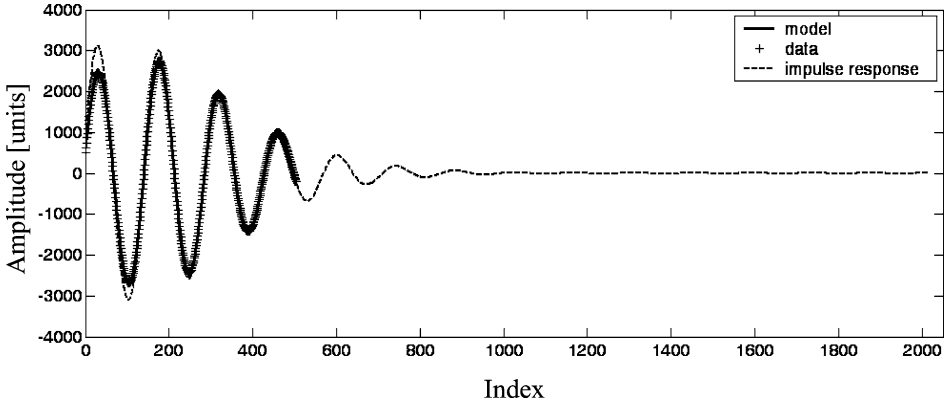


Fig. 5. The data and estimated model for the 3-rd formant interval.

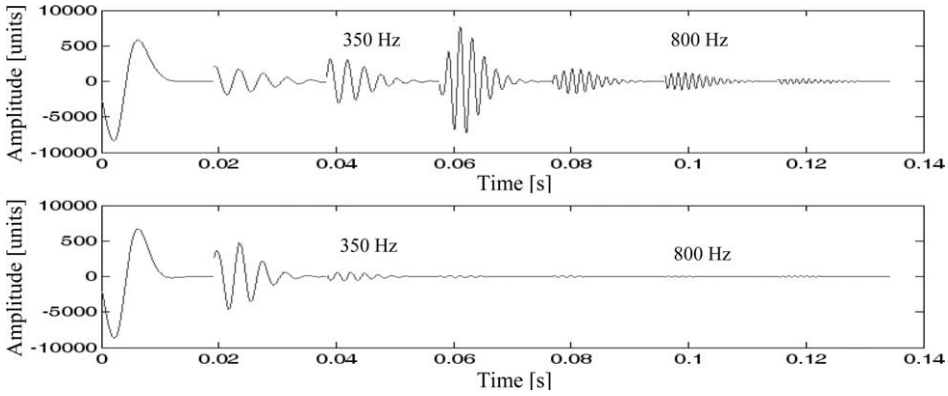


Fig. 6. The vowel /a/ and /i/ formants with frequencies from the bandwidth of 30–1000 Hz (the upper plot – formants of the vowel /a/, the lower plot – formants of the vowel /i/).

of iterations is less than 100 or the estimation error is less than 0.5. We obtained the following estimates $\Lambda = -0.0095, \Omega = 0.044$[rad/sample].

The model and the data for the 3-rd formant interval are shown in Fig. 5.

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 Tables 2 and 3, respectively.

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 Figs. 6 and 7.

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. 8.

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 %.

Table 2
Formant parameters of the vowels /a/

| Formant number $k$ | Frequency [Hz] $f_k$ | Damping [unit] $\lambda_k$ | Amplitude 1 [unit] $A_{1k}$ | Amplitude 2 [unit/s] $A_{2k}$ | Amplitude 3 [unit/s$^2$] $A_{3k}$ | Phase 1 [rad] $\varphi_{1k}$ | Phase 2 [rad] $\varphi_{2k}$ | Phase 3 [rad] $\varphi_{3k}$ |
|---|---|---|---|---|---|---|---|---|
| 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 |

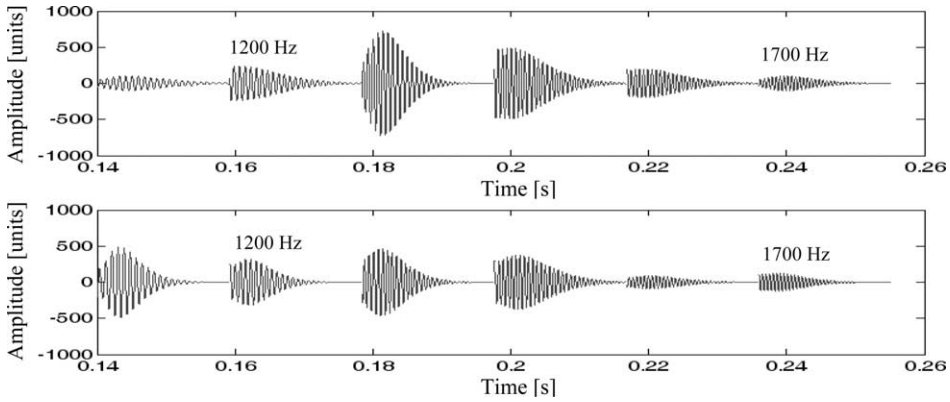


Fig. 7. The vowel /a/ and /i/ formants with frequencies from the bandwidth of 1001–2000 Hz (the upper plot – formants of the vowel /a/, the lower plot – formants of the vowel /i/).

Table 3
Formant parameters of the vowels /i/

| Formant number $k$ | Frequency [Hz] $f_k$ | Damping [unit] $\lambda_k$ | Amplitude 1 [unit] $A_{1k}$ | Amplitude 2 [unit/s] $A_{2k}$ | Amplitude 3 [unit/s$^2$] $A_{3k}$ | Phase 1 [rad] $\varphi_{1k}$ | Phase 2 [rad] $\varphi_{2k}$ | Phase 3 [rad] $\varphi_{3k}$ |
|---|---|---|---|---|---|---|---|---|
| 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 |

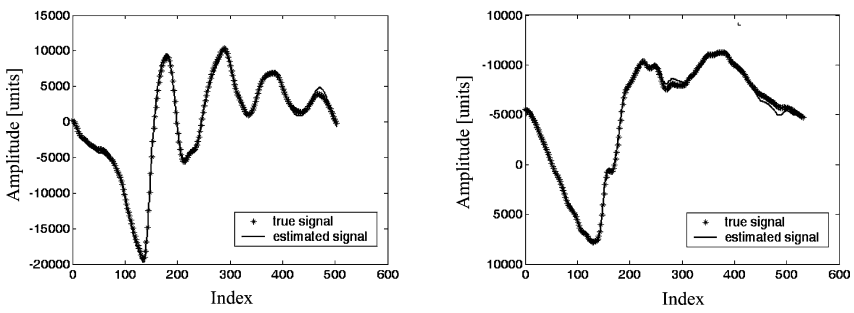


Fig. 8. The true and estimated signal for the 3rd formant interval.

In order to get the system input impulse amplitudes, we calculated the Fourier transform of the diphthong /ai/. This transform was then filtered in the intervals shown in Table 1. 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, we computed the inverse Fourier transform and got signals in the time domain whose length was 22,560 each. We then searched the local maxima of each of these signals and stored them in columns of two matrices corresponding to the vowels /a/ and /i/. The values of these columns are shown in Fig. 9.

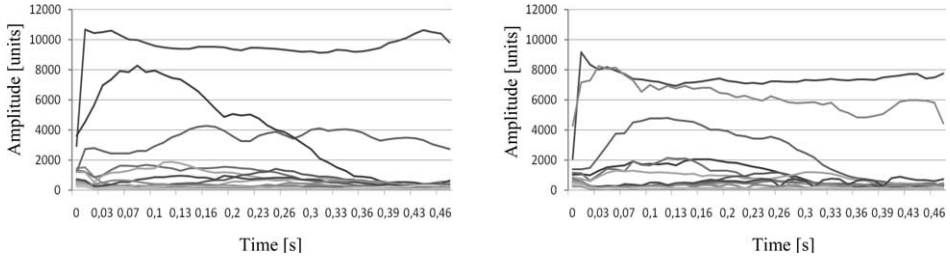


Fig. 9. The input impulse amplitudes (the left figure – for the vowel /a/, the right figure – for the vowel /i/).

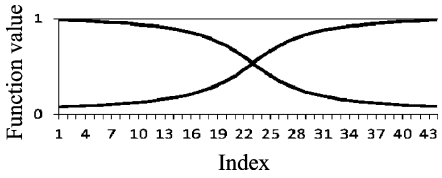


Fig. 10. The values of the arccotangent function arccot($x$) and those of the arctangent function arctan($x$) $+\pi/2$ used to decrease/increase the input impulse amplitudes for the vowels /a/ and /i/.

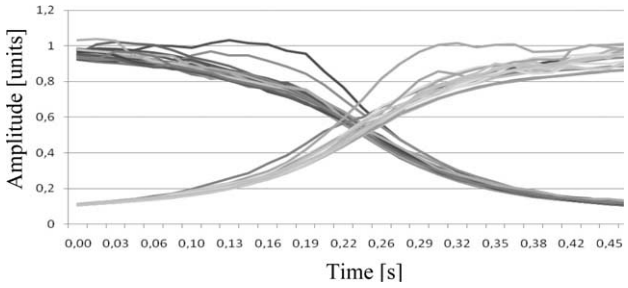


Fig. 11. Input amplitude dynamics.

After analyzing the values of these matrices, we see, that these values are rather large. Therefore we norm them using the following formula:

$$a_{ij} = 1 + (a_{ij} - 1) \cdot c, \tag{51}$$

where $a_{ij}$ is the $(i, j)$th entry of the appropriate matrix, and $c$ is a constant: $c = 0.0001$.

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 $\mathrm{arcctg}\,(x)$ while these corresponding to the vowel /i/ – by the arctangent function (more precisely, by the function $\mathrm{arctg}\,(x)+\pi/2$ (see Fig. 10). The result is shown in Fig. 11. We use multiplication by these functions 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/.

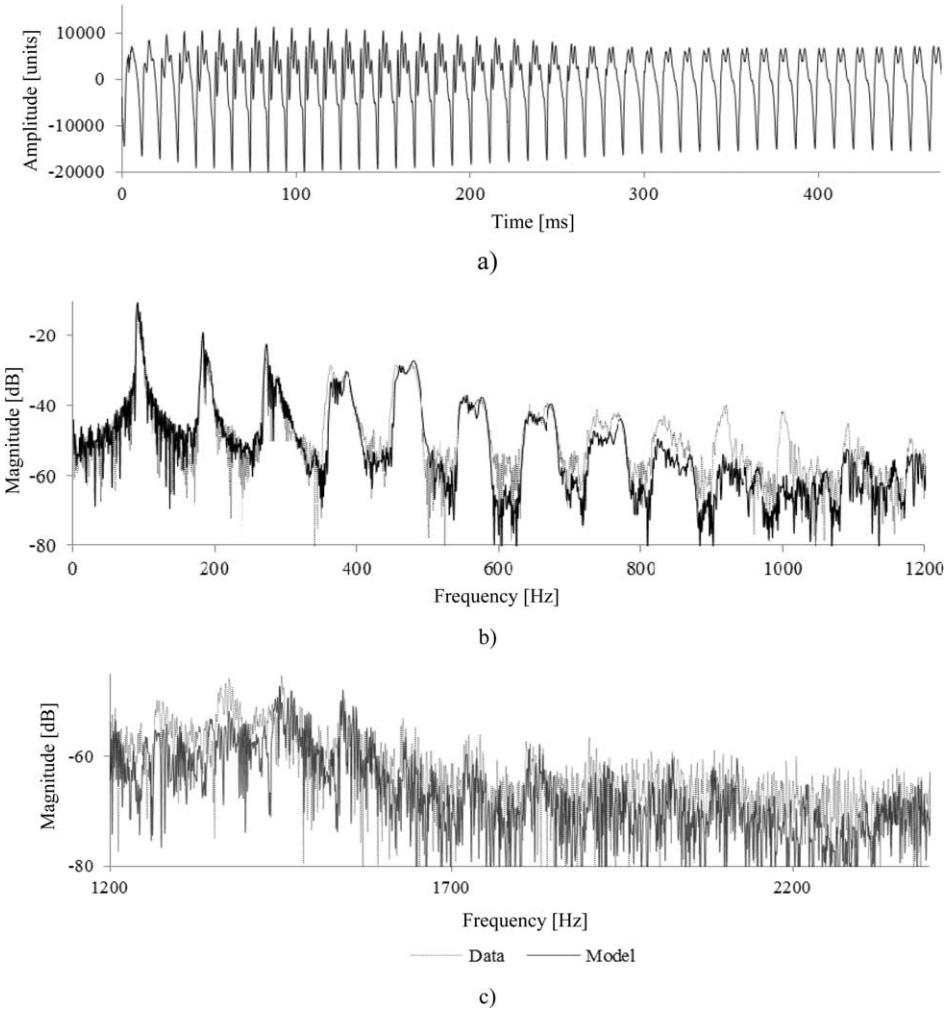


Fig. 12. The Fourier transform of the output process of the synthesyzer "ai": (a) the output signal ; (b) the magnitude response in the range 1–1200 Hz; (c) the magnitude response in the range 1200–2400 Hz.

Now when we have calculated the inputs and impulse response of our MISO system, we obtain the system output using formula (8). In Fig. 12, we present a fragment of the modelled signal (the diphthong /ai/) and its magnitude response.

## 6. Conclusions

1. High modelling quality was achieved due to: (a) the high order of quasipolynomial models, (b) the large number of formants (about 40), (c) separate excitation impulse sequences for each formant.

2. The convoluted system matrix (that is used for parameter estimation) is ill-conditioned, therefore each formant's parameters were estimated separately. We filtered data for each formant band using the inverse Fourier transform. That allowed to estimate all formant parameters sufficiently stable.
3. We used the assumption that the impulse response of the SISO system corresponding to a single formant decays after three fundamental periods. This assumption allowed us to apply Levenberg–Marquardt method to estimate formant parameters from convoluted data.
4. It was almost impossible to distinguish between real and simulated diphthongs in various Lithuanian words with a help of audiotesting. The accuracy of the selected two pitches (one for vowel /a/, and one for vowel /i/) modells was high (about 5%). 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.
5. We think that the same approach can also be applied to other diphthongs. The additional research, however, is needed.
6. Our study shows that it is possible to develop an automatic system that generates a diphthong model using recorded speaker data. A certain programming work should be done in the future.

## References


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.

*Collins Concise English Dictionary* (2009). http://www.thefreedictionary.com/diphthong.

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

Cook, P.R. (2002). *Real Sound Synthesis for Interactive Applications*, Vol. 1. A.K. Peters Ltd, Natick.

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.

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

Fox, A. (2005). *The Structure of German*. Oxford University Press, London.

Fry, D. (1977). *Man as a Talking Animal*. Cambridge University Press, Cambridge.

Garšva, K. (2001). Complex diphthongs ie, uo and their phonological interpretation. A man and a word. *Didactic linguistics*. 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*, 35, 35–38. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.5424.

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.

*Handbook of the International Phonetic Association*. International Phonetic Association (1999), Cambridge.

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

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

Huang, Y., Benesty, J., Chen, J. (2010). *Acoustic MIMO Signal Processing*. Springer, Berlin.

Kajackas, A., Anskaitis, A. (2009). An investigation of the perceptual value of voice frames. *Informatica*, 4(20), 487–498.

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. (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.

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, 164–168.

Lipeika, A. (2010). Optimization of formant feature based speech recognition. *Informatica*, 3(21), 361–374.

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., Lipeikienė, J., Telksnys, L. (2002). Development of isolated word speech recognition system. *Informatica*, 1(13), 37–46.

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

Mannell, R. H. (1998). Formant diphone parameter extraction utilising a labelled single-speaker database. In: *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, Australia. `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.

Martirosian, O., Davel, M. (2008). Acoustic analysis of diphthongs in standard South African English. In: *Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa,* Cape Town, South Africa, pp. 153–157.

Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V. (2009). Investigation of foreign languages models for Lithuanian speech recognition. *Elektronika ir elektrotechnika*, 3, 15–20.

Mobius, B., van Santen, J. (1996). Modelling segmental duration in German text-to-speech synthesis. In: *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, USA, Vol. 4, pp. 2395–2398. `http://www.asel.udel.edu/icslp/cdrom/vol4/652/a652.pdf`.

Norkevičius, G., Raškinis, G. (2008). Modelling phone duration of Lithuanian by classification and regression trees, using very large speech corpus. *Informatica*, 2(19), 271–284. `http://www.onelook.com/?w=phone&ls=a`.

*Oxford dictionaries* (2010). `http://www.askoxford.com/concise_oed/phoneme?view=uk`.

Raškinis, G., Raškinienė, D. (2003b). 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. In: *The Materials of the 4th International Conference on Electrical and Control Technologies ECT-2009*, Technologija, Kaunas, pp. 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. In: *Proceedings of the 5th International Conference on Electrical and Control Technologies ECT-2010*, Kaunas, Lithuania, pp. 117-120.

Rosner, B.S., Pickering, J.B. (1994). *Vowel Perception and Production*. Oxford University Press, London.

Rudžionis, A., Ratkevičius, K., Dumbliauskas, T., Rudžionis, V. (2008). Control of computer and electric devices by voice. *Elektronika ir elektrotechnika*, 6, 11–16.

*SIL International* (2004). `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.

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 (in Russian; republished by Booksurge, USA, 2007).
Šveikauskienė, D. (2005). Graph representation of the syntactic structure of the Lithuanian sentence. *Informatica*, 3(16), 407–418.
Welling, L., Ney, H. (1998). Formant estimation for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(6), 36–48.

**G. Pyž** received her BS degree in mathematics and MS degree in informatics (with honours) from Vilnius Pedagogical University, Lithuania, in 2007 and 2009 respectively. She is currently a PhD student in Vilnius University Institute of Mathematics and Informatics. Her research interests include speech modelling and synthesis.

**V. Šimonytė** graduated from Vilnius University, Lithuania, in 1982 receiving her diploma in applied mathematics (with honours). She received the doctoral degree from Kaunas Polytechnic Institute in 1988. Currently she is inposition of associate profesor at Faculty of Mathematics and Informatics of Vilnius Pedagogical University. Her current research interests include speech modelling and synthesis.

**V. Slivinskas** graduated from Kaunas Polytechnic Institute, Lithuania, in 1967 receiving his diploma in radio engineering. He received the doctoral degree from Kaunas Polytechnic Institute in 1974. V. Slivinskas carried out habilitation procedure in Vilnius Gediminas Technical University in 2009. Currently he is in position of professor at Faculty of Mathematics and Informatics of Vilnius Pedagogical University. His current research interests include signal format modelling and speech synthesis.

## Lietuvių šnekos dvibalsių modeliavimas

Gražina PYŽ, Virginija ŠIMONYTĖ, Vytautas SLIVINSKAS

Straipsnio tikslas yra sukurti lietuvių šnekos dvibalsių modeliavimo metodą. Modeliavimui yra naudojamas formantinis sintezatorius. Antros eilės kvazipolinomas yra pasirinktas kaip formantinis modelis laiko srityje. Daugelio įėjimų ir vieno išėjimo (DĮVI) sistema yra dvibalsio modelis. Ši sistema susideda iš dviejų dalių: pirmoji dalis atitinka pirmąją dvibalsio balsę, o antroji dalis – antrąją balsę. Sistema yra žadinama pusiauperiodinių impulsų seka, naudojant perėjimą iš vienos balsės į kitą. Straipsnyje yra išvestos parametrinės įėjimo–išėjimo lygtys kvazipolinominių formančių atveju, apibrėžta nauja konvoliuotos bazinių signalų matricos sąvoka, išvestos parametrinės minimizavimo funkcionalo formulės konvoliuotiems duomenims. Taip pat straipsnyje yra sukurtas naujas formančių parametrų įvertinimo algoritmas konvoliuotiems duomenims, pagrįstas Levenbergo–Markvarto metodu, ir pateikti šio algoritmo realizavimo žingsniai. Kaip nagrinėjimo pavyzdys, yra pasirinktas lietuvių šnekos dvibalsis /ai/. Šis dvibalsis įrašytas su tokiais parametrais: 48 kHz moduliacijos dažniu, 16 bitų rezoliucija, stereo režimu. Tyrimui naudojami du charakteringi balsių /a/ ir /i/ periodai. Šių balsių DĮVI formantinių modelių parametrų įvertinimui naudojami minėtų periodų vienodai nutolusios atskaitos. Perėjimas tarp balsių /a/ ir /i/ yra gaunamas, keičiant žadinimo impulsų amplitudes pagal arktangento dėsnį. Sukurtajam metodui atliktas audio testavimas bei palygintos realių duomenų ir DĮVI modelio išėjimo Furjė transformacijos. Audiotestas parodė, kad neįmanoma atskirti realaus ir modeliuoto dvibalsių. Palyginus šių dvibalsių amplitudines ir fazines charakteristikas, pastebėti tik labai nežymūs skirtumai.