Developing Models of Lithuanian Speech Vowels and Semivowels

Gražina PYŽ[1]*, Virginija ŠIMONYTĖ[2], Vytautas SLIVINSKAS[2]

[1] *Institute of Mathematics and Informatics, Vilnius University*
Akademijos 4, LT-08663 Vilnius, Lithuania
[2] *Faculty of Science, Mathematics and Technologies*
Lithuanian University of Educational Sciences,
Studentų 39, LT-08106 Vilnius, Lithuania
e-mail: grazina.pyz@mii.vu.lt, virginija.simonyte@leu.lt, vytautas.slivinskas@leu.lt

Received: January 2013; accepted: October 2013

Abstract. Lithuanian vowel and semivowel phoneme modelling framework 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. The synthesized sounds obtained by the harmonic synthesis method are compared with those obtained by the formant method. Application of this modelling framework to all of Lithuanian vowel and semivowel synthesis gives naturally sounding result.

Key words: Lithuanian speech, vowels, semivowels, speech modelling, MISO system, formant, quasipolynomial model, parameter estimation, speech synthesis.

1. Introduction

Lithuanian speech digital processing is a popular area of research by Lithuanian scientists and engineers (see, for example, Kazlauskas, 1999; Lipeikienė and Lipeika, 1998; Maskeliūnas *et al.*, 2009; Tamulevičius *et al.*, 2010). Lithuanian speech synthesis is a part of this area that attracts considerable attention (Kasparaitis, 2001).

There exist two main speech signal synthesis types: concatenative synthesis and formant synthesis (Donovan, 1996; Frolov and Frolov, 2003; Cook, 2002). 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. The synthesized sound is obtained as an output of a linear filter and is described by a mathematical model with a finite number of parameters.

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

[*] Corresponding author.

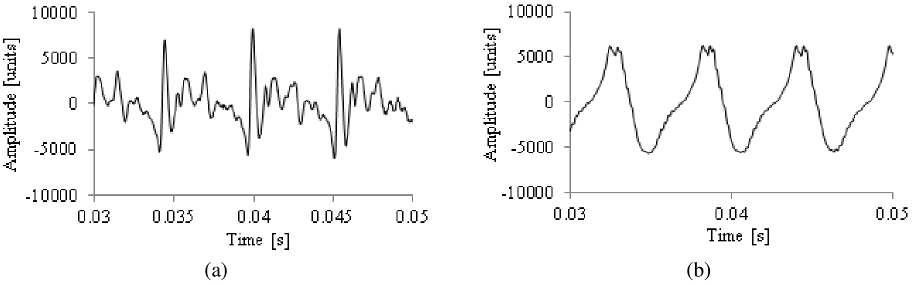

Fig. 1. The periodic character of phonemes: (a) the plot of the vowel /a/, (b) the plot of the semivowel /m/.

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).

Both speech synthesis types attract attention of researchers. Concatenative synthesis of Lithuanian speech was studied in the Kasparaitis papers (1999, 2000, 2005). Methods of quality improvement in concatenative speech synthesis for Polish language were considered in Janicki (2004). What concerns Lithuanian speech, formant synthesis has not yet attracted much attention of researchers. The most known Lithuanian speech synthesizer is based on concatenative synthesis (Balbonas, 2009; Kasparaitis, 2001). The practical implementation can be seen in (Garsiai.lt). Problems related to developing of Lithuanian speech formant synthesizers are considered in Ringys and Slivinskas (2009), Ringys and Slivinskas (2010), Pyž *et al.* (2011). A phoneme synthesis method based on the phoneme signal expansion into harmonics is proposed in our paper (Pyž *et al.*, 2012).

In this paper, we present an improved version of the harmonic method and compare it with an alternative formant method. The paper is organized as follows. The vowel and semivowel phoneme signals decomposition into harmonics is given in Section 2. The vowel and semivowel phoneme signals decomposition into formants is described in Section 3. The vowel and semivowel phoneme model is considered in Section 4. Section 5 deals with selection of the phoneme representative period. Determining of the inputs is described in Section 6. The modelling accuracy is analysed in Section 7.

2. Vowel and Semivowel Phoneme Signals Decomposition into Harmonics

Our goal is to get mathematical models of the analysed phoneme, which could be used as a base of phoneme synthesizer. In general case, the character of vowel and semivowel signals is periodic (see Fig. 1).

One can see from Fig. 1 that a phoneme signal has a rather complex form. It is very difficult to find such a model that fits the phoneme signal well. We use the approach of expanding of a complex signal into the sum of simpler signals. We suggest expanding the signal into the sum of a finite number of components which we call the phoneme signal harmonics (similarly as harmonics of a periodic signal in Fourier series theory).

Table 1
The frequency band partition into subbands.

Subband number	Subband
1	$[0.5f_0, 1.5f_0)$
2	$[1.5f_0, 2.5f_0)$
...	
K	$[((K-1)+0.5)f_0, (K+0.5)f_0)$

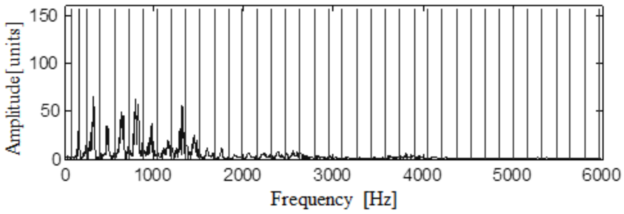

Fig. 2. The magnitude response of a vowel and its partition into subbands.

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

$$s(n) = h_1(n) + h_2(n) + \cdots + h_K(n), \quad n = 1, \dots, N, \tag{1}$$

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

In order to decompose the phoneme signal into harmonics, we have to estimate the fundamental frequency of this signal. Below we present the recurrent algorithm of estimating of the fundamental frequency and calculating harmonics. At first, we calculate the magnitude response of the whole vowel phoneme signal. Let f_0 be an initial estimate of the fundamental frequency. We take this estimate equal to the frequency of the first pick of the spectrum. Then we partition the frequency band 0–6000 Hz into the subbands shown in Table 1.

K is the largest integer number for which the inequality $(K+0.5)f_0 \leqslant 6000$ holds

$$K = [6000/f_0 - 0.5], \tag{2}$$

where $[x]$ stands for the integer part of a real number x. 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. 2.

In each subband, we determine the highest amplitudes $a_1, a_2, \dots, a_K$, and find the frequencies corresponding to those amplitudes: $g_1, g_2, \dots, g_K$. At first glance, these frequencies look like the formants, this, however, is not true in general case. Then we compare the frequency sequences $f_0, 2f_0, \dots, Kf_0$ and $g_1, g_2, \dots, g_K$. Our goal is to find such an f_0 that minimizes the sum of the distances between the frequencies:

$$S_0 = \sum_{k=1}^{K} |kf_0 - g_k|. \tag{3}$$

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) K (K is defined by (2)).
3. The values of the harmonic frequencies $g_1, g_2, \dots, g_K$.

The steps of the algorithm are listed below.

```
Step 1. Compute the sum of the distances $S_0 = \sum_{k=1}^{K} |kf_0 - g_k|$.
Step 2. Set $\triangle = 1$
($\triangle$ 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 $f_{new} = f_0 + \triangle$.
Step 4. Compute the sum of the distances $S_{new} = \sum_{k=1}^{K} |kf_{new} - g_k|$.
Step 5. If $S_{new} < S_0$ then
            $f_0 = f_{new}$,    $S_0 = S_{new}$
        else
            $f_{new} = f_0 - \triangle$,    $S_{new} = \sum_{k=1}^{K} |kf_{new} - g_k|$
            if $S_{new} < S_0$ then
                $f_0 = f_{new}$,    $S_0 = S_{new}$
            else
                $\triangle = \triangle/2$.
Step 6. If $\triangle < 0.01$ then
                Go to Step 7
        else
                Go to Step 3
Step 7. END
```

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

A block diagram of the algorithm presented above is shown in Fig. 3.

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 kf_0 and g_k can not decrease all the time; at a certain time it will start to increase.

After obtaining the optimal value $\tilde{f_0}$, we can decompose the phoneme signal into K harmonics. For this purpose, we make a new frequency band partition into subbands according to Table 1 (with $f_0 = \tilde{f_0}$).

We introduce an auxiliary function $g_k(m)$ defined as follows:

$$g_k(m) = \begin{cases} FFT(s(m)), & m \in [((k-1)+0.5)\tilde{f_0}, (k+0.5)\tilde{f_0}], \\ 0, & m \notin [((k-1)+0.5)\tilde{f_0}, (k+0.5)\tilde{f_0}], \end{cases} \tag{4}$$

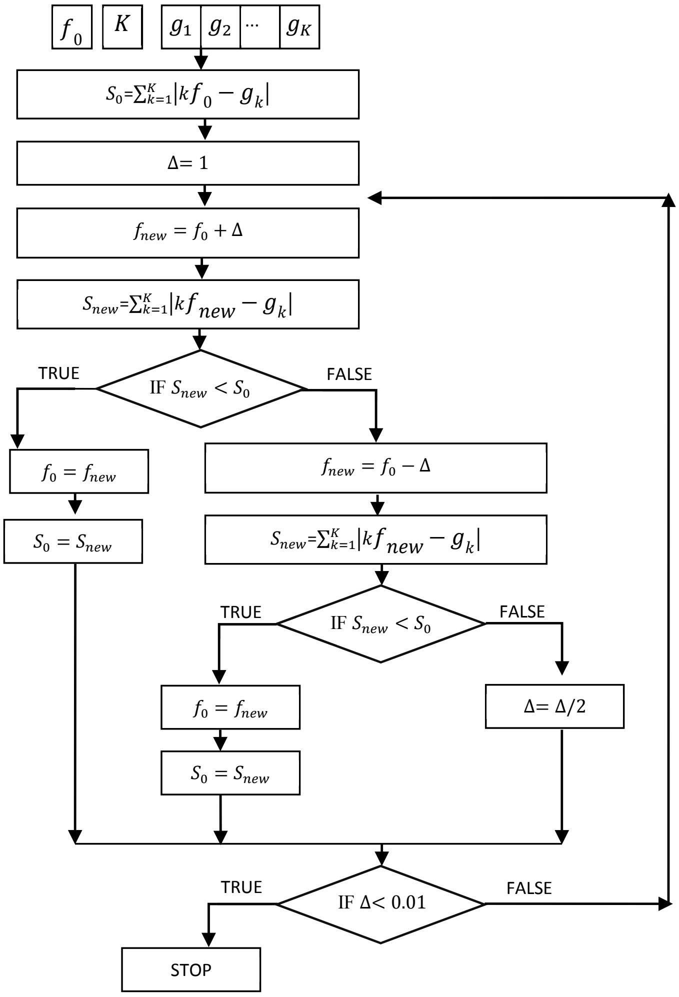

Fig. 3. A block diagram of the fundamental frequency refining algorithm.

where $k=1,\dots,K$, and compute its inverse Fourier transform

$$\tilde{h}_k(n)=\left(\frac{1}{N}\right)\sum_{m=1}^{N} g_k(m)e^{(2\pi i)(n-1)\frac{m-1}{N}}, \tag{5}$$

$n=1,\dots,N$, i – imaginary unit.

The obtained signal $\tilde{h}_k$ is the k-th harmonic of the phoneme signal. The first three harmonics of the female vowel /a:˜/ are shown in Fig. 4.

We see from Fig. 4 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.

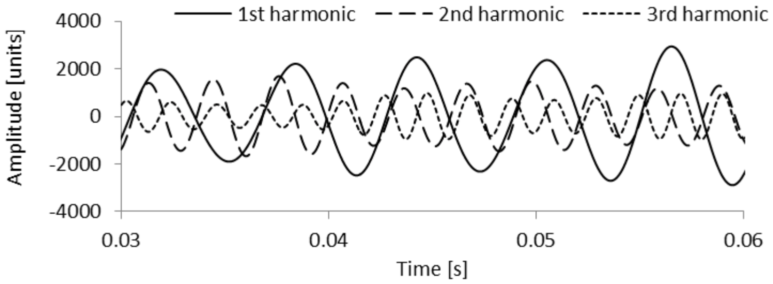

Fig. 4. The first three harmonics of the phoneme /a:˜/ (as in the word ãčiū).

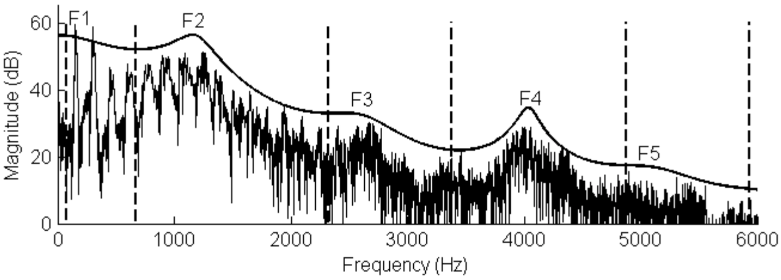

Fig. 5. The plots of the spectrum of the phoneme /a:˜/.

3. Vowel and Semivowel Phoneme Signals Decomposition into Formants

A formant (formant frequency) is defined in a usual way as amaximum 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 (Cook, 2002). In the current paper, we use the Linear Predictive Coding (LPC) method (Markel and Gray, 1976) as a formant extraction tool. We need to partition the frequency band0-6000 Hz into the subbands where each band corresponds to one formant. We execute the partition 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. We propose to add the neighbouring harmonics (calculated in Section 2) corresponding to a selected formant of spectrum. We call the obtained signal a *formant*. Joining of the harmonic frequencies of the phoneme spectrum into groups corresponding to the particular formant frequencies is shown in Fig. 5.

The first three formant components of the female vowel /a:˜/ are shown in Fig. 6.

We see from Figs. 4 and 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.

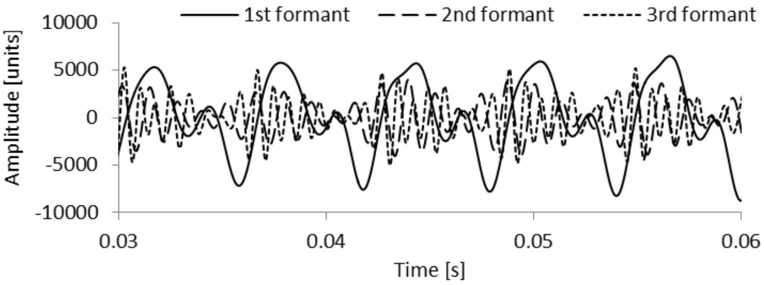

Fig. 6. The first three formant components of the phoneme /a:˜/.

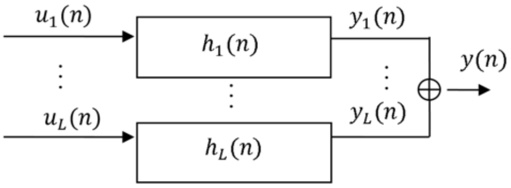

Fig. 7. A MISO system for vowel phoneme modelling.

4. The Vowel and Semivowel Phoneme Model

We have to expand a phoneme signal into components. It is natural to choose the parallel connection model where each component is modelled separately. We propose to use for modelling such a discrete time linear stationary system with L inputs and a single output (L is the number of harmonics (in harmonic synthesis method) or formants (in formant synthesis method)). 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. 7, where $\{u_l(n)\}$, $\{h_l(n)\}$, $\{y_l(n)\}$ ($n = 0, 1, 2, \ldots$) are the sequences of the input, impulse response and output of the l-th single-input and single-output (SISO) system, and $\{y(n)\}$ is the output sequence of the multiple-input and single-output (MISO) system. 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 formant we get a signal of a certain form. This signal is similar to a quasipolynomial that is the product of a sinusoid and polynomial. We consider such a MISO system in which each SISO system (sometimes called a channel) impulse response is described by a third degree quasipolynomial

$$\begin{aligned} h_l(n) &= e^{\Delta t \lambda_l t}\big(a_{l1}\sin(2\pi f_l n\Delta t + \varphi_{l1}) + a_{l2}t\sin(2\pi f_l n\Delta t + \varphi_{l2}) \\ &\quad + a_{l3}t^2\sin(2\pi f_l n\Delta t + \varphi_{l3}) + a_{l4}t^3\sin(2\pi f_l n\Delta t + \varphi_{l4})\big) \end{aligned} \tag{6}$$

where $\lambda_l < 0$ is the damping factor, f_l – the frequency, a_{li} ($i = 1, \ldots, 4$) – the amplitudes, φ_{li} ($i = 1, \ldots, 4$) – the phases, $\Delta t = 1/f_s$ (f_s – the sampling frequency). In order to get

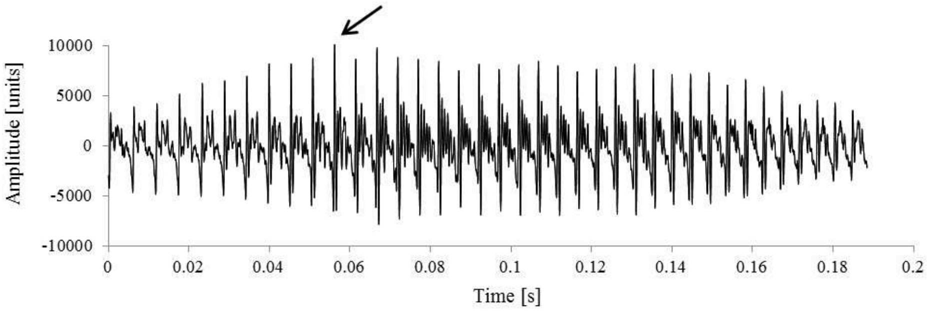

Fig. 8. Selecting of the vowel phoneme pitch with the highest amplitude.

smoother formant (harmonic) signals we set the amplitude $a_{l1} = 0$. Computations show that a third degree quasipolynomial model is a good trade-off between the result quality and model complexity.

5. 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 we consider only a single period for each phoneme. Such a period is usually called a pitch.

Let

$$\mathbf{y}^M = \left[y(0), y(1), y(2), \ldots, y(M-1)\right]^T \tag{7}$$

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 4).In order to automate this selection procedure, we choose the amplitude size as the selection criteria, i.e. the pitch with the highest amplitude is selected as a representative one. We are looking for a representative period that is within the 60% of the phoneme signal samples (i.e. we reject 20% of the samples in the signal start part and 20% in the signal end part). Figure 8 shows a vowel phoneme signal made of several pitches where the pitch with the highest amplitude is marked with an arrow.

We use the representative period to compute the parameters of the MISO system impulse response components (6). The start point of the representative pitch (the point t_{start} in Fig. 9) is selected in the following way:

1. The phoneme signal is filtered in the low-pass filter with the $2.5 f_0$ 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 coincides with the phoneme signal periods. The filtered signal is shown in Fig. 9.
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

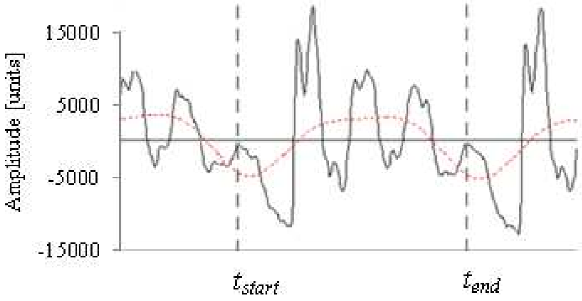

Fig. 9. Selecting of the start and end points of the representative pitch.

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 we look for a point 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 we fail to find such a point, we look for 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. 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. 9) is as follows. We select the pitch start point and find the point that is at the distance of $T_0 = 1/f_0$ from this start point. With a help of the found point we determine the pitch end point in the same way as we described above (with the start point).

The parameters of the impulse responses are estimated using Levenberg–Marquardt method. A step-by-step algorithm of this method for a second degree quasipolynomial is described in the paper (Pyž *et al.*, 2011). After a simple modification the algorithm can be easily adapted for a third degree quasipolynomial.

6. Determination of the Inputs

A MISO system proposed in Section 4 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. 10), then a signal with equal periods is obtained in the system output; those periods are the same as the representative period.

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 of such impulses is presented in this section.

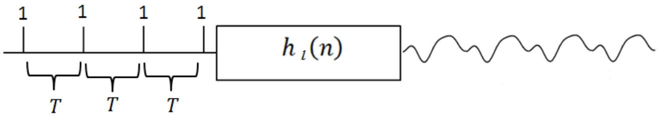

Fig. 10. A SISO system with the unit impulse inputs.

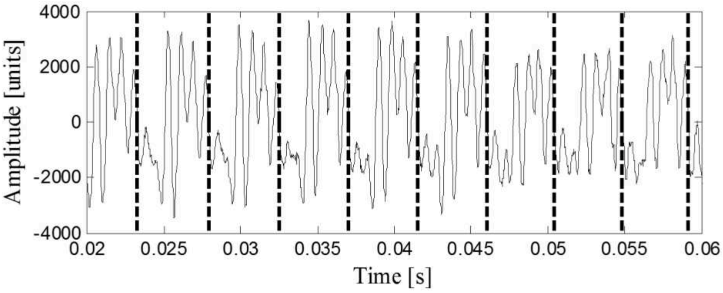

Fig. 11. A part of the phoneme /a/ signal divided into periods.

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 4). An example of such segmentation is shown in Fig. 11.

After segmenting a signal into periods, we determine the start point of each period. We then divide the first harmonic or formant (harmonic – in harmonic synthesis method, formant – in formant synthesis method) component into time intervals using the determined points. In each of these intervals, we find the maximum points. The ordinates of these points are stored in a vector that is the first column of the matrix that we call the input matrix. Analogously we segment the second component into the same time intervals. We then find again the maximum point in each of these intervals. 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: we select a row that corresponds to the representative pitch and divide all the values of that column by the value at intersection of this column and the selected row. After the norming procedure is completed, we obtain the ratios of the representative period amplitude and amplitudes of the all periods. 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. 12 (the harmonic synthesis method case).

From Fig. 12 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 significantly reduce.

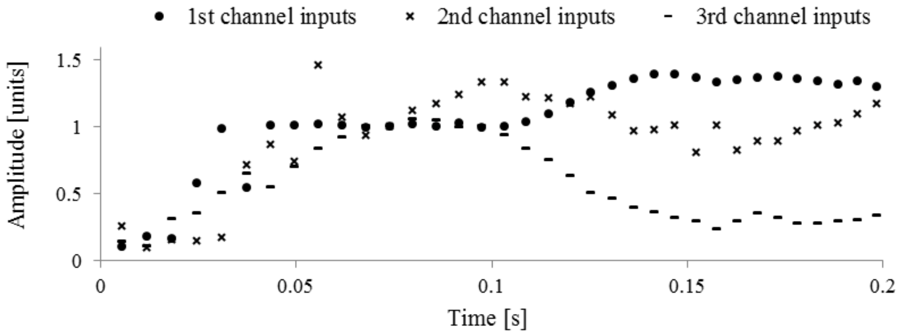

Fig. 12. The inputs of the first three channels of a MISO system.

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

$$T = [T_1, T_2, \dots, T_P]. \qquad (8)$$

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

7. Experimental Results

Lithuanian language phonemes (sounds) have been studied in Girdenis (1995). A. Girdenis listed 58 phonemes in this work. 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 of 4 additional phonemes (Kasparaitis, 2005). All the phonemes mentioned above along with a pause make a list of 92 units.

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 Lithuanian consonants “j”, “l”, “m”, “n”, “r”, “v” are called semivowels as they have both vowel and semivowel features. Lithuanian language has nineteen pure semivovel phonemes. 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”.

We carried out the modelling for all the vowel and semivowel phonemes using 50 utterances by female and 50 utterances by male. The list of Lithuanian words used in the experiment is presented in Appendix A. In order to estimate the model quality we calculated the average spectrum. The comparison of the spectra of the true phoneme /a/ signal and its models is shown in Fig. 13.

Figure 13 shows that the obtained spectra almost coincide. The audio test revealed that the differences have no significant influence to the sound intelligibility and quality.

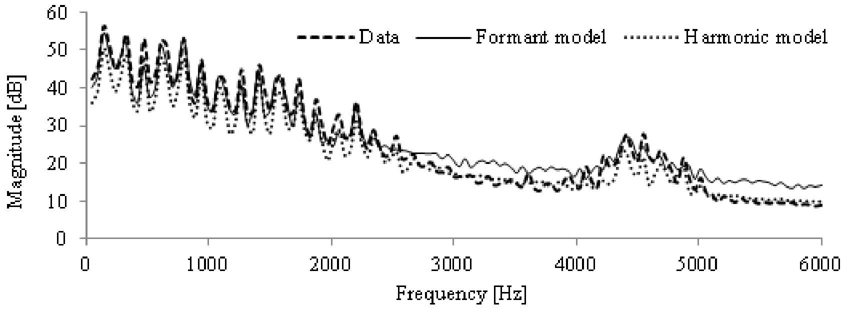

Fig. 13. The spectra of the true phoneme /a/ signal and its models.

Table 2
The average RMSE and its confidence intervalsfor 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%	

The average root-mean-square-error (RMSE) of the estimated signal spectrum and its confidence intervals for each of the 28 vowel and 19 semivowel phonemes are presented in Tables 2 and 3. The RMSE is computed by the following formula:

Table 3
The average RMSE and its confidence intervalsfor 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%	

$$RMSE = 100\% \times \sqrt{\frac{\sum_{q=1}^{Q} (S_q - \hat{S}_q)^2}{Q}}, \tag{9}$$

where S_q is the q-th value of the spectrum of the true phoneme, $\hat{S}_q$ – the q-th value of the spectrum of the modelled phoneme, Q is the number of the spectrum values of true/modelled phoneme. The confidence intervals are stated at the 95% confidence level.

The graphical representation of the average RMSE shown in Tables 2 and 3 is presented in Figs. 14 and 15.

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.

8. Conclusions

The main contribution of this paper is avowel and semivowel phoneme synthesis framework that is based ona vowel and semivowel phoneme mathematical model and an au-

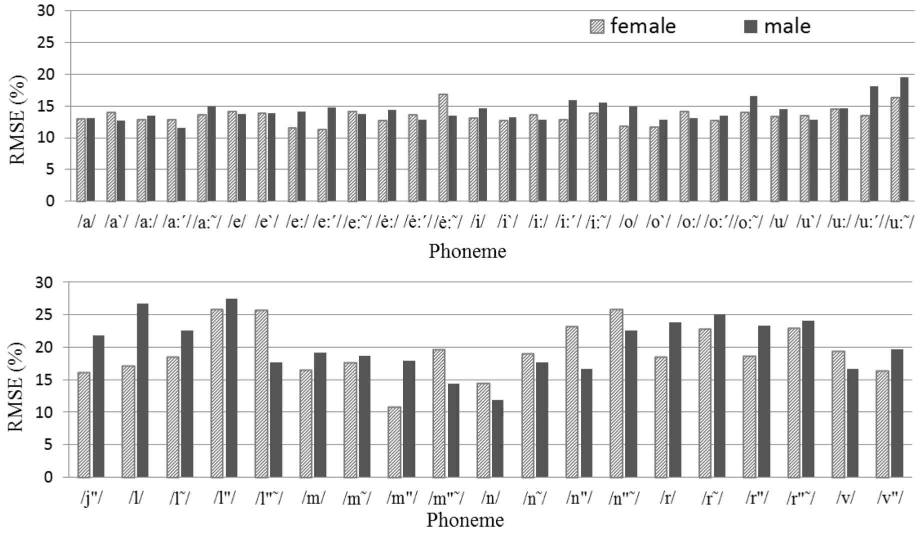

Fig. 14. The average RMSE for the estimated signal spectrum (formant method case): *the upper plot* – vowel phonemes, *the lower plot* – semivowel phonemes.

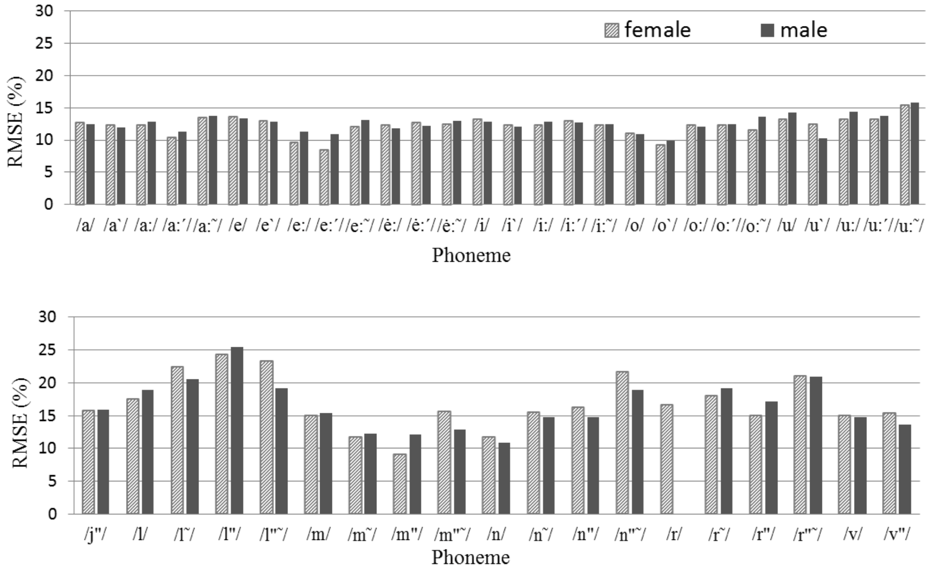

Fig. 15. The average RMSE for the estimated signal spectrum (harmonic method case): *the upper plot* – vowel phonemes, *the lower plot* – semivowel phonemes.

tomatic procedure of estimation of the vowel phoneme fundamental frequency and input determining.

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 the synthesized signal sounds more naturally, we use impulses of different amplitudes and periods as inputs instead of the unit impulses with a constant period.

A new fundamental frequency refining algorithm is proposed.

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

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 sufficiently small estimation errors (13% for vowels and 18% for semivowels)and audio test show that the proposed framework gives sufficiently good vowel and semivowel synthesis quality.

Appendix

A list of Lithuanian phonemes (originally created by Girdenis (1995) and appended by Kasparaitis (2005)) with the examples of their usage (presented by the authors of this paper).

	Phoneme	Description	Example
1	/a/	The short unstressed vowel /a/	m**a**mà ‘mother’
2	/a`/	The short stressed vowel /a/	lazd**à** ‘stick’
3	/a:/	The long unstressed vowel /a/	dr**ą**sà ‘courage’
4	/a:´/	The long vowel /a/ stressed with the falling accent	k**á**rdas ‘sword’
5	/a:˜/	The long vowel /a/ stressed with the rising accent	**ã**čiū ‘thank you’
6	/e/	The short unstressed vowel /e/	m**e**dãlis ‘medal’
7	/e`/	The short stressed vowel /e/	sug**è**sti ‘turn bad’, ‘get out of order’
8	/e:/	The long unstressed vowel /e/	gr**ę**žinỹs ‘well’, ‘borehole’
9	/e:´/	The long vowel /e/ stressed with the falling accent	**é**rkė ‘mite’
10	/e:˜/	The long vowel /e/ stressed with the rising accent	gyv**ẽ**nimas ‘life’
11	/ė:/	The long unstressed vowel /ė/	k**ė**d˜ė ‘chair’
12	/ė:´/	The long vowel /ė/ stressed with the falling accent	up´**ė**takis ‘trout’
13	/ė:˜/	The long vowel /ė/ stressed with the rising accent	gėl˜**ė** ‘flower’
14	/i/	The short unstressed vowel /i/	l**i**gà ‘disease’, ‘illness’
15	/i`/	The short stressed vowel /i/	k**ì**škis ‘rabbit’
16	/i:/	The long unstressed vowel /i/	t**y**là ‘silence’, ‘quiet’
17	/i:´/	The long vowel /i/ stressed with the falling accent	r**ý**tas ‘morning’
18	/i:˜/	The long vowel /i/ stressed with the rising accent	arkl**ỹ**s ‘horse’
19	/o/	The short unstressed vowel /o/	f**o**toaparãtas ‘camera’
20	/o`/	The short stressed vowel /o/	ch**ò**ras ‘choir’
21	/o:/	The long unstressed vowel /o/	k**o**vótojas ‘fighter’, ‘warrior’
22	/o:´/	The long vowel /o/ stressed with the falling accent	š**ó**nas ‘side’
23	/o:˜/	The long vowel /o/ stressed with the rising accent	Ad**õ**mas ‘Adam’
24	/u/	The short unstressed vowel /u/	k**u**ltūrà ‘culture’
25	/u`/	The short stressed vowel /u/	**ù**pė ‘river’
26	/u:/	The long unstressed vowel /u/	k**ū**rinỹs ‘work’, ‘piece’
27	/u:´/	The long vowel /u/ stressed with the falling accent	l´**ū**pa ‘lip’
28	/u:˜/	The long vowel /u/ stressed with the rising accent	m˜**ū**šis ‘battle’
29	/j"/	The soft consonant /j/ (in Lithuanian, ‘j’ is always soft)	**j**´ūra ‘sea’ **j**ìs ‘he’

Continued in next page

	Phoneme	Description	Example
30	/l/	The consonant /l/	vá**l**sas ‘waltz’
31	/l˜/	The stressed consonant /l/	vi**l̃**kas ‘wolf’
32	/l"/	The soft consonant /l/	va**l**ià ‘will’
33	/l"˜/	The soft stressed consonant /l/	gu**l̃**ti ‘to go to bed’, ‘to lie down’
34	/m/	The consonant /m/	ã**m**atas ‘handicraft’
35	/m˜/	The stressed consonant /m/	li**m̃**palas ‘adhesive ’
36	/m"/	The soft consonant /m/	s**m**ẽgenys ‘brain’
37	/m"˜/	The soft stressed consonant /m/	ka**m̃**štis ‘cork’
38	/n/	The consonant /n/	**n**ãmas ‘house’
39	/n˜/	The stressed consonant /n/	i**ñ**karas ‘anchor’
40	/n"/	The soft consonant /n/	**n**èšti ‘to carry along’
41	/n"˜/	The soft stressed consonant /n/	le**ñ**ktis ‘to bend’, ‘to bow’
42	/r/	The consonant /r/	**r**ãtas ‘wheel’ be**r**žas ‘birch’
43	/r˜/	The stressed consonant /r/	ga**r̃**sas ‘sound’
44	/r"/	The soft consonant /r/	k**r**iáušė ‘pear’
45	/r"˜/	The soft stressed consonant /r/	ki**r̃**tis ‘stress’, ‘blow’
46	/v/	The consonant /v/	**v**óras ‘spider’
47	/v"/	The soft consonant /v/	**v**iáuksėti ‘to yelp’

References

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

Cook, P.R. (2002). *Real Sound Synthesis for Interactive Applications*, Vol. 1. A.K. Peters Ltd., Natick, 263 p.

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

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

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

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

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.

Kazlauskas, K. (1999). Noisy speech intelligibility enhancement. *Informatica*, 2(10), 171–188.

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

Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V. (2009). Investigation of foreign languages models for Lithuanian speech recognition. *Electronics and Electrical Engineering*, 3(91), 15–20.

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

Oxford Dictionary of English (2010). Oxford University Press, London.

Pyž, G., Šimonytė, V., Slivinskas, V. (2011). Modellingof Lithuanian speech diphthongs. *Informatica*, 3(22), 411–434.

Pyž, G., Šimonytė, V., Slivinskas, V. (2012). Lithuanian speech synthesizing by computer using additive synthesis. *Electronics and Electrical Engineering*, 8(18), 77–80.

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)*, pp. 117–120.

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.

G. Pyž received her BS degree in mathematics and MS degree in informatics (with honours) from Lithuanian University of Educational Sciences (former Vilnius Pedagogical University), Lithuania, in 2007 and 2009 respectively. She received the doctoral degree from Vilnius University Institute of Mathematics and Informatics in 2013. Currently she works in this institution. 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 got the doctoral degree from Kaunas University of Technology (former Kaunas Polytechnic Institute) in 1988. At present she is in position of associate professor at Faculty of Science and Technology of Lithuanian University of Educational Sciences. Her current research interests include speech modelling and synthesis.

V. Slivinskas graduated from Kaunas University of Technology, Lithuania, in 1967 receiving his diploma in radio engineering. He received the doctoral degree from Kaunas University of Technology in 1974. V. Slivinskas carried out habilitation procedure at Vilnius Gediminas Technical University in 2009. At present he is in position of professor at Faculty of Science and Technology of Lithuanian University of Educational Sciences. His current research interests include signal formant modelling and speech synthesis.

Lietuviškos šnekos balsių modelių kūrimas

Gražina PYŽ, Virginija ŠIMONYTĖ, Vytautas SLIVINSKAS

Lietuviškų balsių ir pusbalsių fonemų modeliavimo sistema yra pateikta. Naudojant šią sistemą, fonemos signalas yra aprašomas kaip tiesinės dinaminės sistemos su daugeliu įėjimų ir daugeliu išėjimų signalas išėjime. Pastaroji dinaminė sistema yra lygiagretus vieno įėjimo ir vieno išėjimo sistemų junginys. Straipsnyje yra pasiūlyti du sintezės metodai: harmoninis ir formantinis. Sintezuoti garsai, gauti harmoninės sintezės metodu, yra lyginami su garsais, gautais formantiniu metodu. Taikant šią modeliavimo sistemą lietuviškų balsių ir pusbalsių sintezei, gaunamos natūraliai skambančios fonemos.