Lithuanian Speech Synthesis by Computer using Additive Synthesis

G. Pyz[1], V. Simonyte[2], V. Slivinskas[2]
[1]*Recognition Processes Department, VU Institute of Mathematics and Informatics, Akademijos St. 4, LT-08663 Vilnius, Lithuania, phone: +370 5 2660380, e-mail: grazinute123@gmail.com*
[2]*Department of Informatics, Faculty of Mathematics and Informatics, Lithuanian University of Educational Sciences, Studentų St. 39–415, LT-08106 Vilnius, Lithuania, phone: +370 5 2751796, grazinute123@gmail.com*

***Abstract*—We present a new Lithuanian speech phoneme synthesis method based on the principle of additive synthesis in this paper. An assumption is made that phoneme models consist of the sum of harmonics which could be generated by properly chosen formant synthesizer parameters. In order to estimate the synthesizer parameters, we use the real sound signals that are expanded into harmonics by the inverse fast Fourier transform method. The harmonic synthesizer parameters (amplitudes, damping factor, and phases) are estimated by Levenberg-Marquardt method. We present an example of the synthesized female vowel /a/ and compare it with the true sound signal.**

***Index Terms*—Speech synthesis, Lithuanian language, phoneme, Fourier series, additive synthesis method.**

I. INTRODUCTION

Much attention in Lithuania is given to processing Lithuanian speech and applying the results for speech recognition, animation, analysis and synthesis. The most efforts are devoted to speech recognition. The latest work is concentrated in the fields of developing new methods for speech recognition feature quality measurement, control of computer and electric devices by voice, etc.

Speech animation problems attract also attention of Lithuanian researchers. One of such problems is Lithuanian phoneme visualization. A methodology of such visualization is proposed in [1]. Some researchers try to use Wiener class systems for speech signal prediction.

An important class of speech processing problems is speech synthesis. Speech synthesis methods can be divided into two main groups: parametric synthesis methods and concatenation synthesis methods. In parametric speech synthesis, a speech signal is represented by a finite number of parameters. In concatenation synthesis, a sound is created with a help of a predefined vocabulary of the initial synthesis elements. The most known work in the field of text-to-speech synthesis of Lithuanian language is [2]. Concatenation synthesis is used in this work. The practical implementation of the methodology proposed in [2] can be seen in [3]. The main concatenation synthesis problem is the size of the memory for storing the vocabulary. The synthesized speech quality does not achieve the natural speech quality since glitches occur on the concatenation boundaries.

A subgroup of parametric synthesis methods is the so-called formant synthesis methods. These methods are based on the decomposition of a speech signal into formants (signals described by quasipolynomial models) [4]–[7]. A vowel formant synthesizer has been developed in [4]. The authors of [4] proposed to synthesize Lithuanian speech vowels in the two frequency ranges: the low frequency range (1-900 Hz) and the high frequency range (900-2400 Hz). The examples of two synthesized vowels /a/ and /i/ have been presented. This algorithm gave satisfactory results (the naturally sounding vowel models), although certain parameters were not selected automatically. The authors of [5] did not divide the frequency range into two parts and excited each formant separately. In order to improve parameter estimates, the optimization procedure has been introduced in [5]. The purpose of this optimization is to reduce errors due to data convolution. The amplitudes of actual formants have been used as inputs, and the distance between the inputs was equal to the main pitch periods of the original sound. The models of the excitation sequences and fundamental frequency dynamics, however, have not been developed.

Formant diphthong model has been developed in [6]. For developing of this model, the transition between the diphthong vowels has been described using the arctangent function. An algorithm for the estimation of parameters has been derived from convoluted data, and it was applied for synthesizing of the diphthong /ai/. The same methodology as in [6] has been applied for joining of vowels, diphthongs and semivowels [7].

The synthesized sounds obtained by methods of [4]–[7] have sufficiently natural sounding. The synthesizing procedures described in [4]–[7], however, have a disadvantage – it is difficult to determine the formant ranges and decompose the signal into formants as there is uncertainty related with hidden and merged formants.

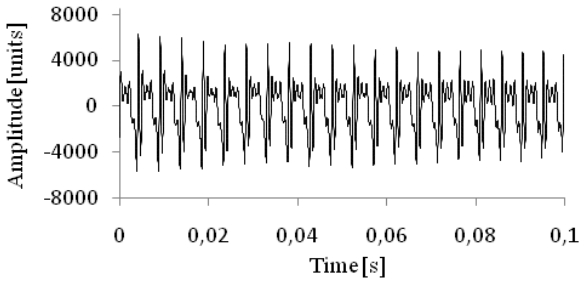
In this paper, as an alternative to the expanding of a signal into formants, we use the expanding of a signal into

Manuscript received March 15, 2012; accepted May 14, 2012.

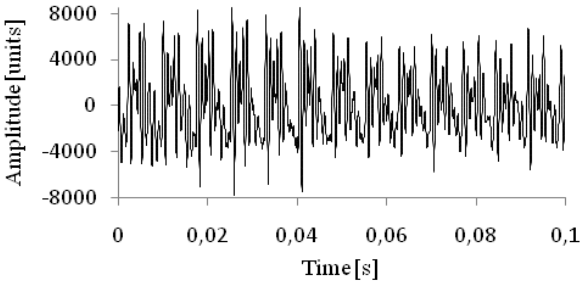
harmonics. Since the distance between the harmonics frequencies is known, then the expanding of a signal into harmonics can be done without uncertainty.

II. ADDITIVE SYNTHESIS OF A SPEECH SIGNAL

A short-time speech signal can be regarded as periodic. The periodic character can be seen in Fig. 1 where a female vowel /a/ and male vowel /a/ are shown.



a)



b)

Fig. 1. Female and male vowels /a/ (a) – female, (b) – male.

Mathematically, a periodic signal $y(t)$ satisfies the following relationship

$$y(t) = y(t+T),\ T > 0. \tag{1}$$

A periodic function can be expanded into a Fourier series

$$y(t) = \sum_{k=1}^{\infty} a_k \sin(2\pi k f_0 t + \varphi_k), \tag{2}$$

where $f_0 = 1/T$ is the signal fundamental frequency, a_k is the amplitude of the *k*-th harmonic, φ_k is the phase of the *k*-th harmonic.

A finite number of harmonics K is used to synthesize speech sounds using (2) because very high harmonics do not almost affect the speech signal sound. Then the relationship (2) is changed as follows

$$y(t) = \sum_{k=1}^{K} a_k \sin(2\pi k f_0 t + \varphi_k)\,. \tag{3}$$

One can encounter various transitions in speech signals (from one phoneme to another, from the phoneme beginning to the phoneme end, amplitude jumps in a stressed syllable, etc.). The phases are also important in order to avoid phase distortions in transitions. The sound synthesized using (3) has strong synthetic shade. Therefore in order to get a natural sounding, it is assumed that the harmonic amplitudes and the fundamental frequency are functions of time. Then the relationship (3) turns into the following

$$y(t) = \sum_{k=1}^{K} a_k(t) \sin(2\pi k f_0(t) t + \varphi_k). \tag{4}$$

The speech sound synthesis by (4) is carried out in two steps: 1) K harmonics are synthesized, and 2) the synthesized harmonics $h_i(t)$ are summed.

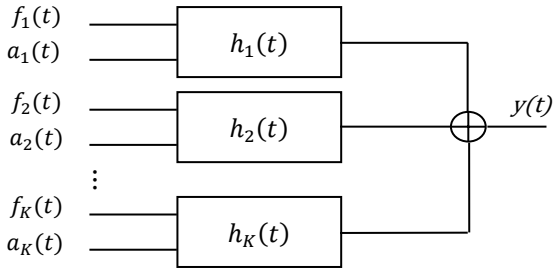


Fig. 2. The speech synthesis scheme using (4).

III. GENERATING OF SPEECH SIGNAL HARMONICS USING THE FORMANT SYNTHESIZER

It is not difficult to show that a formant synthesizer [6] can generate sinusoid-type signals by properly choosing parameters. Therefore we suggest to use a formant synthesizer for harmonic synthesis, i. e. to use a linear system with unit pulse inputs whose impulse response is a third order quasipolynomial

$$\begin{gathered} h_i(t) = a_{i1} e^{-\lambda_i t} \sin(2\pi f_i t + \varphi_{i1}) + \\ + a_{i2} t f_d e^{-\lambda_i t} \sin(2\pi f_i t + \varphi_{i2}) + a_{i3} t^2 f_d^2 e^{-\lambda_i t} \sin(2\pi f_i t + \\ + \varphi_{i3}) + a_{i4} t^3 f_d^3 e^{-\lambda_i t} \sin(2\pi f_i t + \varphi_{i4}), \end{gathered} \tag{5}$$

where i is the harmonic number, t – continuous time, f_d – the sampling frequency, f_i – the harmonic frequency, λ_i – the damping factor, a_{i1}, a_{i2}, a_{i3}, a_{i4} – the amplitudes, φ_{i1}, φ_{i2}, φ_{i3}, φ_{i4} – the phases. The computations are carried out not with continuous functions but with the sequences obtained by sampling continuous-time signals. We therefore use the following discrete-time state space model [8]:

$$\begin{cases} x(k+1) = \boldsymbol{F}x(k) + \boldsymbol{G}u(k), \\ y(k) = \boldsymbol{H}x(k), \end{cases} \tag{6}$$

where **F** is a block diagonal matrix made of K Jordan blocks:

$$F_i = \begin{bmatrix} \begin{matrix} a & -b \\ b & a \end{matrix} & \begin{matrix} a & -b \\ b & a \end{matrix} & \begin{matrix} a/2 & -b/2 \\ b/2 & a/2 \end{matrix} & \begin{matrix} a/6 & -b/6 \\ b/6 & -a/6 \end{matrix} \\ 0 & \begin{matrix} a & -b \\ b & a \end{matrix} & \begin{matrix} a & -b \\ b & a \end{matrix} & \begin{matrix} a/2 & -b/2 \\ b/2 & -a/2 \end{matrix} \\ 0 & 0 & \begin{matrix} a & -b \\ b & a \end{matrix} & \begin{matrix} a & -b \\ b & a \end{matrix} \\ 0 & 0 & 0 & \begin{matrix} a & -b \\ b & a \end{matrix} \end{bmatrix}, \tag{7}$$

$$\begin{cases} a = a_i = e^{\lambda_i \Delta t} \cos(2\pi f_i \Delta t)\,, \\ b = b_i = e^{\lambda_i \Delta t} \sin(2\pi f_i \Delta t)\,, \end{cases} \tag{8}$$

where K – the total number of harmonics, **G** – the block

diagonal matrix with K column vectors $g = [0,0,0,0,0,0,1,1]^T$ on the main diagonal:

$$H = [H_1, H_2, \dots, H_K], \tag{9}$$

$$H_i = [3\alpha_{i4}\beta_{i4}, 3\alpha_{i4}\gamma_{i4}, \alpha_{i3}\beta_{i3}, \alpha_{i3}\gamma_{i3}, 0.5\alpha_{i2}\beta_{i2}, \times \\ \times 0.5\alpha_{i2}\gamma_{i2}, 0.5\alpha_{i1}\beta_{i1}, 0.5\alpha_{i1}\gamma_{i1}], \tag{10}$$

$$\beta_{ij} = \sin(\varphi_{ij}) - cos(\varphi_{ij}), \tag{11}$$

$$\gamma_{ij} = \sin(\varphi_{ij}) + +cos(\varphi_{ij}), \tag{12}$$

where i=1,2,…,K, j=1,…,4.

In order to use (6) for harmonic synthesis, one has to estimate all the parameters of the system.

IV. ESTIMATION OF PARAMETERS OF THE FORMANT HARMONIC SYNTHESIZER FROM REAL DATA

To estimate the model parameters, we first expand the speech signal into harmonics using rectangular filters that are implemented by the inverse Fourier transform. The magnitude response of the female vowel /a/ is shown in Fig. 3. and the signal harmonics in Fig. 4.

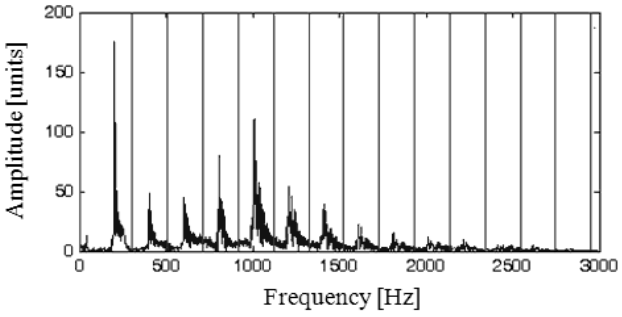


Fig. 3. The amplitude response of the female vowel /a/.

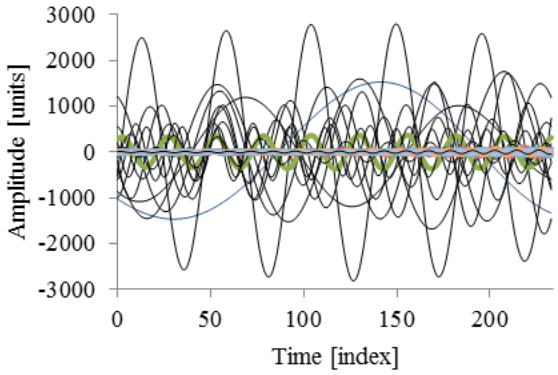


Fig. 4. The harmonics of the female vowel /a/.

In order to estimate the formant filter parameters, we select a "representative" pitch (Fig. 5). For this purpose, we select the minimum point 1, then go up until the first maximum 2, and go down to the first minimum 3.

A 'step-by-step' Levenberg-Marquardt type algorithm described in [8] is used for parameter estimation from convoluted data. The estimation of parameters of the harmonic synthesis system is carried out in parallel for each harmonic data. The seventh speech signal harmonic and the seventh harmonic of the synthesizer output signal are shown in Fig. 6.

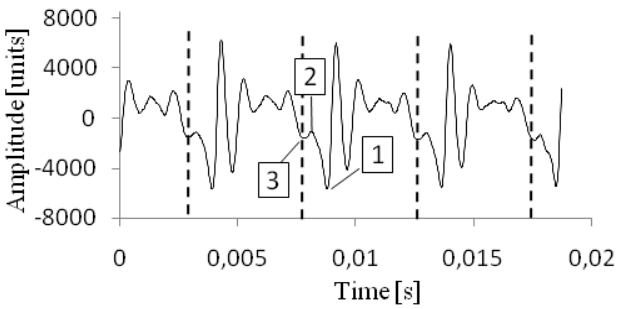


Fig. 5. Selecting of the representative pitch.

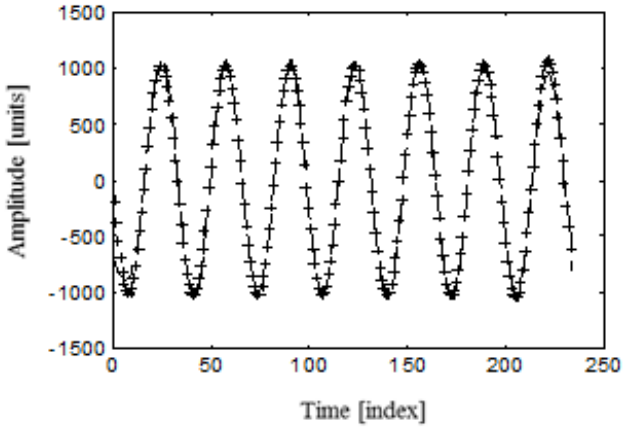


Fig. 6. The seventh speech signal harmonic and the seventh harmonic of the synthesizer output signal ('+' – the data, the solid line – the synthesizer output).

In order to check the synthesizer accuracy, we carried out the following experiment: three unit pulses with the period $T = 1/f_0$ were sent to the synthesizer input and the output signal was compared with the "representative" pitch. The true and model signals are shown in Fig. 7.

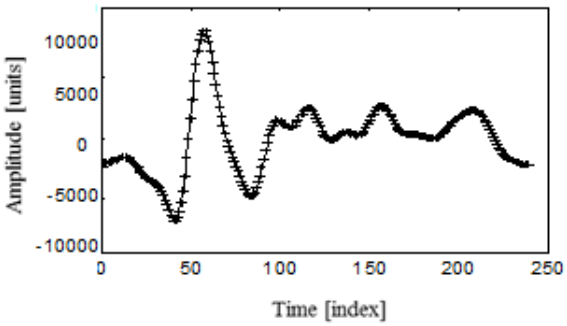


Fig. 7. The true and model signals ('+' – true signal, the solid line – model).

V. CONCLUSIONS

A new harmonic synthesis method belonging to the additive synthesis class is proposed as an alternative to the formant synthesizer. The experimental results show that the synthesized sounds of this method are sufficiently natural, pleasantly sounding. The third-order polynomial models are used for amplitudes and periods of excitation pulse sequence dynamics modelling.

Although the synthesizer model seems to be of a high order and has an excessive number of parameters, it performs an important function providing naturalness to the synthesized speech. When necessary, the model can be

reduced, for example, by combining a few adjacent harmonics into one formant.

REFERENCES

[1] E. Ivanovas, D. Navakauskas, "Peculiarities of Wiener Class Systems and their Exploitation for Speech Signal Prediction", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)*, no. 5, pp. 107–110, 2011.

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

[3] *Lithuanian speech synthesis.* [Online]. Available: http://www.garsiai.lt

[4] T. Ringys, V. Slivinskas, "Formant modelling of Lithuanian language vowel natural sounding", in *Proc. of. the Materials of the 4th International Conference on Electrical and Control Technologies ECT-2009*, 2009, pp. 5–8.

[5] T. Ringys, V. Slivinskas, "Lithuanian language vowel formant modelling using multiple input and single output linear dynamic system with multiple poles", in *Proc. of. the Materials of the 4th International Conference on Electrical and Control Technologies ECT-2010*, 2010, pp. 117–120.

[6] G. Pyž, V. Šimonytė, V. Slivinskas, "Modelling of Lithuanian Speech Diphthongs", *Informatica*, no. 3, pp. 411–434, 2011.

[7] G. Pyž, V. Šimonytė, V. Slivinskas, "Joining of Vowel and Semivowel Models in Lithuanian Speech Formant-based Synthesizer", in *Proc. of. the Materials of the 4th International Conference on Electrical and Control Technologies ECT-2011*, 2011, pp. 114–119.

[8] V. Slivinskas, V. Šimonytė, *Minimal Realization and Formant Analysis of Dynamic Systems and Signals*. Mokslas. Vilnius, 1990, p. 168. (in Russian, republished by Booksurge, USA, 2007).