ARCHIVES OF ACOUSTICS
Vol. **42**, No. 3, pp. 375–383 (2017)

DOI: 10.1515/aoa-2017-0039

Voiceless Stop Consonant Modelling and Synthesis Framework Based on MISO Dynamic System

Gražina KORVEL[(1)], Božena KOSTEK[(2)]

[(1)] *Institute of Mathematics and Informatics*
Vilnius University
4 Akademijos Str., Vilnius LT-08663, Lithuania; e-mail: grazina.korvel@mii.vu.lt

[(2)] *Audio Acoustics Laboratory*
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
G. Narutowicza 11/12, 80-233 Gdańsk, Poland; e-mail: bokostek@audioakustyka.org

(*received February 1, 2017; accepted April 11, 2017*)

A voiceless stop consonant phoneme modelling and synthesis framework based on a phoneme modelling in low-frequency range and high-frequency range separately is proposed. The phoneme signal is decomposed into the sums of simpler basic components and described as the output of a linear multiple-input and single-output (MISO) system. The impulse response of each channel is a third order quasi-polynomial. Using this framework, the limit between the frequency ranges is determined. A new limit point searching three-step algorithm is given in this paper. Within this framework, the input of the low-frequency component is equal to one, and the impulse response generates the whole component. The high-frequency component appears when the system is excited by semi-periodic impulses. The filter impulse response of this component model is single period and decays after three periods. Application of the proposed modelling framework for the voiceless stop consonant phoneme has shown that the quality of the model is sufficiently good.

Keywords: speech synthesis; consonant phonemes; phoneme modelling framework; MISO system.

1. Introduction

In recent years, speech technology has made rapid advances in many areas such as automatic speech recognition (ASR), automatic audio-visual speech recognition (AVSR), automatic transcription, building meaningful multimodal speech corpora, etc. Numerous examples of national speech corpora other than English exist (e.g. AGH Corpora; Brocki, Marasek, 2015; Igras *et al.*, 2013; Jadczyk, Ziółko, 2015; Johannessen *et al.*, 2007; Korzinek *et al.*, 2011; Oostdijk, 2000; Pinnis, Auziņa, 2010; Pinnis *et al.*, 2014; Upadhyaya *et al.*, 2015; Stănescu *et al.*, 2012), but in most cases they are devoted to build a material for speech recognition tasks. The common feature of such corpora is a careful analysis of design criteria and search for a relevant speech material. Also, there exist websites,e.g. VoxForge which were set up to collect transcribed speech for use with Open Source Speech Recognition Engines. Though, many challenges such as poor input signal quality, noise and echo disturbance, ambiguity and the use of non-standard phraseology remain, resulting in reducing the recognition rate and the performance of speech recognition systems (Czyzewski *et al.*, 2017). Thus, even though the problem of speech data collecting and analyzing is not new, there are still ongoing research studies on several aspects.

Also, speech synthesis has generated wide interest in speech processing for decades. The dominating speech synthesis technique is unit-selection synthesis (Zen *et al.*, 2009). Many recent studies have focused on using Hidden Markov Model (HMM) in synthesizing speech. A general overview of speech synthesis based on this method is given in the paper by Tokuda *et al.* (2013). Demenko *et al.* (2010) present a study on adapting the open-source software, called BOSS (The Bonn Open Synthesis System), which was originally designed for generating German speech utilizing a concatenative speech synthesis to the Polish language. For that purpose Polish speech corpus based on various databases was created and later evaluated (Demenko *et al.*, 2010; SAMPA, 2005; SAMPA Polish, 2005). As pointed out by the authors of that paper, creating a versatile speech synthesis system is not an obvious task as such a system depends on gathering not only a specific task-oriented speech material, but should be enhanced by co-articulatory effects, enabling to create expressive speech as well (Demenko *et al.*, 2010). It is also interesting that the analysis of dynamic spectral properties of formants may lead to a significant reduction of information carried by speech signal (Gardzielewska, Preis, 2007).

A voice source modelling method based on predicting the time domain glottal flow waveform using a DNN (Deep Neural Network) is described in very recent sources (Raitio *et al.*, 2014). Tamulevičius and Kaukënas (2016) apply Autoregressive model parameter estimation technique for modelling of semivowels. Contrarily, much less attention has been paid to formant speech synthesis. The main reason is that the synthesized speech quality does not achieve the natural speech quality yet (Sasirekha, Chandra, 2012; Tabet, Boughazi, 2011). Formant synthesizers have advantages against the concatenative ones. The speech produced by them can sufficiently be intelligible even at high speed (Tabet, Boughazi, 2011). They can control prosody aspects of the synthesized speech. Still, in order to reduce synthetic sounding, there is a need to develop new mathematical models for speech sounds.

There are about 200 different vowels in the world's languages and more than 600 different consonants (Ladefoged, Disner, 2012). It should be pointed out, that vowel or vowel-consonant modelling is a better exploited subject. Therefore, in this paper the main focus is given to the consonants. The development of consonant models is a classic problem in speech synthesis. The signals of consonant phonemes are more difficult than those of vowels and semivowels. For example, no previous study has considered Lithuanian consonant phoneme models. Most studies in Lithuanian consonant phonemes have only been carried out in the speech recognition area. A system for discrimination of fricative consonants and sonants is proposed in the paper (Driaunys *et al.*, 2012). The work (Raškinis, Dereškeviciutë, 2007) describes an investigation of spectral properties of the voiceless velar stop consonant /k/ of Lithuanian. The phonology of Polish was described in many sources (e.g. Jassem, 2003; Gussmann, 2007; Oliver, Szklanny, 2006), but interestingly also by Labarre (2011). He pointed out that in terms of consonants, one can distinguish 36 contrastive consonant phonemes in Polish (Labarre, 2011). The goal of his study was to show differences between Polish and American English phonology. The study was carried out at the University of Washington by the author having Polish ancestry. In the study of Krynicki (2006) some contrasting aspects of Polish and English phonetics were shown and adequate examples of such were recalled. The acoustic part of the AGH AVSR consists of a variety of speech scenarios, including phonetically balanced 4.5 h subcorpus recorded in an anechoic chamber, which may be useful for extracting material for carrying out evaluation tests (AGH Corpora; Żelasko *et al.*, 2016). The phonetical statistics were collected from several Polish corpora (Ziółko *et al.*, 2009). A corpus of spoken Polish was used to collect statistic values of real language and evaluated to be applied in an automatic speech recognition and speaker identification systems. This feature could be used in phoneme parametrization and modelling (Ziółko, Ziółko, 2011).

A search of world literature revealed few studies which deal with vowel or consonant-vowel modelling (Birkholz, 2013; Stevens, 1993). Mostly, speech organs producing sounds of the given language are considered in these papers. In the current research sound is described in terms of acoustical properties, i.e. signal characteristics are considered. For this purpose, we describe the signal as the output of MISO (multiple-input and single-output) system. The usage of the liner system for speech synthesis is proposed in the paper (Ringys, Slivinskas, 2010). This solution requires estimation of the filter parameters and inputs.

The object of this research is voiceless stop consonant phonemes. The phonemes /b/, /b'/, /d/, /d'/, /g/, /g'/, /k/, /k'/, /p/, /p'/, /t/, /t'/ are called stop consonants because the air in the vocal tract is stopped at some period. We can divide those phonemes into two sets: voiced and voiceless sounds (Domagała, 1994; Krynicki, 2006). The difference between these sets lies in the action of the vocal folds. For phonemes /b/, /b'/, /d/, /d'/, /g/, /g'/, the vocal folds vibrate while saying these sounds. Therefore, they are called voiced sounds. Meanwhile for voiceless phonemes /k/, /k'/, /p/, /p'/, /t/, /t'/ the vocal folds are apart.

The main purpose of the investigations reported here is to propose a new voiceless stop consonant phoneme modelling and synthesis framework. The synthesis technique presented in this paper enables one to develop phoneme models. The proposed models can be used for developing a formant speech synthesizer which does not use any recorded sounds. These models can also be adapted to other similar problems, for example treating language disorders, speech recognition, helping with pronunciation and learning foreign languages.

The paper starts with introducing the proposed phoneme mathematical model. It then pass to the modelling framework with the main focus on signal dividing into components into low and high-frequency ranges. Then, the paper presents the results of the experiments. Conclusions are presented in the last section.

2. Phoneme mathematical model

The goal of the research is to obtain the mathematical model of the analyzed phoneme. Generally, a phoneme signal has a quite complicated form. It is proposed to expand this signal into the sum of components (formants). Each of these components is responsible for a certain frequency band and is treated as the output of MISO system channel. The diagram of such a system is shown in Fig. 1, where: K – number of components, K_1 – number of low-frequency components, $\{u(n)\}$, $\{h(n)\}$, $\{y(n)\}$ are the sequences of the input, impulse response and output, respectively.

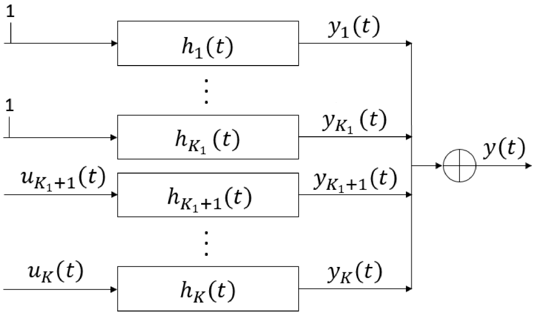

Fig. 1. Multiple channel synthesis scheme.

The expansion of the signal into a sum of formants is needed to satisfy the criteria of the minimal model i.e. the number of formants and the order of formant needs to be as minimal as possible. The impulse response of such a system is described as a third order quasi-polynomial:

$$\begin{aligned} h(t) = {} & \mathrm{e}^{\lambda t}(a_1 \sin(2\pi f t + \varphi_1) \\ & + a_2 t \sin(2\pi f t + \varphi_2) \\ & + a_3 t^2 \sin(2\pi f t + \varphi_3) \\ & + a_4 t^3 \sin(2\pi f t + \varphi_4)), \qquad (1) \end{aligned}$$

where $t \in R^+ \cup \{0\}$, $\lambda > 0$ – the damping factor, f – the frequency, a_k – amplitude, $\varphi_k(-\pi \le \varphi_k < \pi)$ – phase. Computations show (see Pyž *et al.*, 2014) that a third degree quasi-polynomial is a good trade-off between the resulted quality and the model complexity.

The modelling of components consists of two steps, the first of which is the impulse response parameter estimation and the second refers to the determination of the exciting input impulse periods and amplitudes. The parameters of the impulse responses are estimated using the Levenberg-Marquardt method. A step-by-step algorithm of this method for a second-degree quasi-polynomial is described in an earlier paper of one of the authors of this study (Pyž *et al.*, 2011). In order to obtain more natural sounding of the synthesized speech, it is important to use not only high-order models but complex input sequence scenarios as well. A procedure of determining inputs is presented in the more recent paper by Pyž *et al.* (2014).

3. Modelling framework

The analyzing of stop consonant phonemes shows that high frequencies generate sound of the phoneme, contrarily, low frequencies retain timbre of the speaker. Therefore, in this research, it is proposed to divide the phoneme signal into two parts and model it in the high-frequency range and low-frequency ranges separately. In order to divide a phoneme into two parts, it is necessary to set the limit between those frequency ranges. For this purpose, the three-step algorithm is given below:

1) The y-coordinate of the highest point on the given curve is estimated. This value is marked as *max* (see Fig. 2).
2) The point where the line $y = \max/3$ crosses the y-axis is determined. This point is marked as *cross point*.
3) From the *cross point*, we will go down until we reach a minimum. Such a point will be a *limit point*.

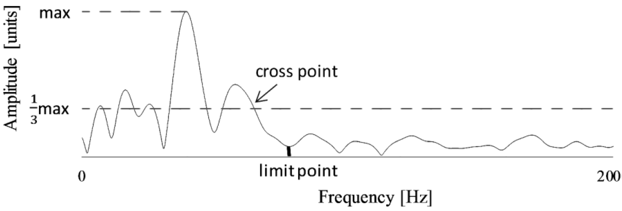

Fig. 2. The magnitude response of the phoneme /k/.

Figure 2 shows the graphical representation of limit point searching. Note that the frequency spectrum from the range [0, 2000] Hz is considered.

3.1. *Signal dividing into components in low-frequency range*

First of all, the signal is filtered with a filter from the bandwidth of [1, f_{limit}] Hz. The second step is to divide the signal into components, each of which contains a single harmonic with inharmonics. In order to determine the partition points, the second order derivative of the spectrum function is computed. The local minima of the derivative are considered as partition points. The length of adjacent periods of consonant phoneme signal (in contrast to vowel and semivowel phoneme) differs slightly from each other. We can consider this signal as quasi-periodic signal in noise. As a result, after dividing the spectrum into components using determined partition points, we obtain some signals which hold inharmonics but do not hold a harmonic. These signals are insignificant. Therefore, it is necessary to reject the points which are adjacent to each other. For this purpose, the near point rejection algorithm is proposed.

Input:

1) $p_1, p_2, \ldots, p_{L_1}$ – the initial partition points,
2) L_1 – number of the initial partition points,
3) d – the allowed minimal distance between points (this value depends on the speaker's fundamental frequency).

Output:

1) $P_1, P_2, \ldots, P_{L_2}$ – the partition points,
2) L_2 – number of the partition points,

A pseudocode of this algorithm is shown below:

```
Set the first partition point to the first
value of the initial partition point list
Loop through each value in the initial
partition point list
    If the distance between this value and the
    last value from the partition point list
    is bigger than the allowed minimal distance
    between points, then set this value to the
    partition point list.
End loop
```

It is worth emphasizing that there are not lower and upper limits to calculate partition points. In order to use the proposed algorithm, we have to determine the allowed minimal distance between points. We assume that each component should have a harmonic. It can be done if the allowed distance is not less than half fundamental frequency (f_0). We set that $d = f_0/2$. Estimation of the fundamental frequency is an active topic of research. Currently there exist many fundamental frequency estimation methods (Dziubiński, Kostek, 2005). An example was proposed by Pyž *et al.* (2014).

A block diagram of this algorithm is presented in Fig. 3.

After applying the algorithm shown in Fig. 3, the near points will be rejected and the number of frequency bands will be equal to K_1 ($K_1 = L_2 - 1$). The frequency band is divided into subbands:

$$g_k(m) = \begin{cases} \mathrm{FT}(m), & m \in [P_k,\, P_{k+1}], \\ 0, & m \notin [P_k,\, P_{k+1}], \end{cases} \tag{2}$$

where FT(m) – Fourier transform of the phoneme signal $s(n)$, $k = 1, \ldots, K_1$.

The component of the phoneme is calculated using the inverse Fourier transform in the corresponding frequency band:

$$\widetilde{h}_k(n) = \left(\frac{1}{N}\right) \sum_{m=1}^{N} g_k(m) \mathrm{e}^{(2\pi i)(n-1)\frac{m-1}{N}}, \tag{3}$$

where N – phoneme length ($n = 1, \ldots, N$), i – imaginary unit.

After implementation of Eq. (3), we obtain K_1 signals of the N point length. These signals are used for parameters of the impulse responses Eq. (1) estimation.

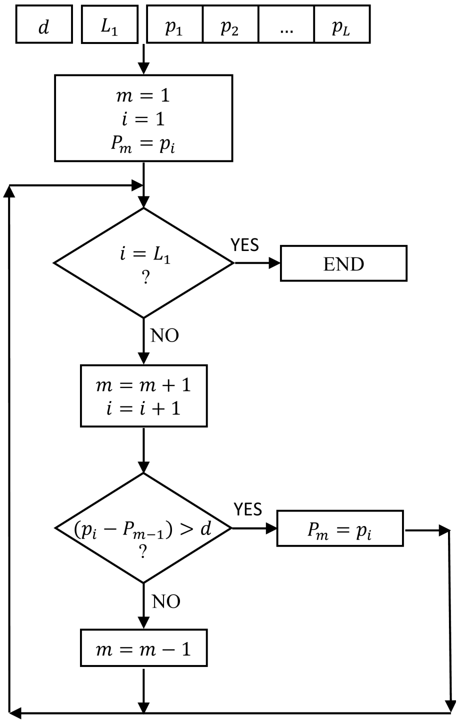

Fig. 3. A block diagram of the near points rejection algorithm.

3.2. Signal dividing into components in the high-frequency range

The signal is filtered with a filter with the bandwidth of [f_{limit}, 8000] Hz. Low frequencies are attenuated and the signal gains the periodic character after voiceless stop consonant filtering in the high-frequency range. Therefore, only a single period is considered. The conditions of the period selection are as follows:

1) The first sample of the period is as close as possible to zero.
2) The energy of the beginning of the period is larger than that of the end.

Such a period is called a representative period and is used for the parameter estimation. The method that allows one to select representative period automatically was given by Pyž *et al.* (2014). The magnitude response of the representative period is calculated. The procedure of determining the partition points is as follows:

1) The first peak of the magnitude response is chosen.
2) The frequency corresponding to this minimum is the first partition point.
3) The second point is obtained analogously, i.e. the second peak of the magnitude response is chosen and then the algorithm proceeds to the right from the peak until the nearest local minimum.

After this procedure, we get K_2 frequency bands. In each of these bands, the inverse Fourier transform is performed. Respectively, K_2 signals are obtained. The length of these signals is equal to the length of the selected period. For each of them, the parameters are estimated.

4. Experimental results

An utterance of the voiceless stop consonant /p'/ is considered in this Section. Its duration is 0.013 s. This consonant was recorded as wav audio file format with the following parameters: PCM 44.1 kHz, 16 bit, stereo. The signal consists of 603 samples and is shown in Fig. 4.

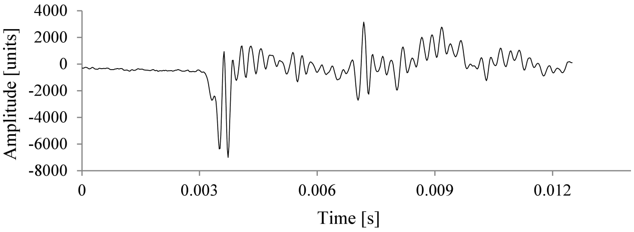

Fig. 4. The oscillogram of voiceless stop consonant /p'/.

First, the magnitude response of this signal is calculated and the limit point between high and low frequencies is determined. After applying the three-step algorithm described in Sec. 3, we get that limit point which is equal to 930 Hz. Then, the signal is filtered with a filter from the bandwidth of 1–930 Hz and the magnitude response of this signal is calculated. The frequency bands are selected as shown in Table 1. After dividing the magnitude response into frequency bands, five intervals are determined. In each of these intervals, the inverse Fourier transform is carried out. As a result, five signals of length 0.013 s are obtained. These signals are shown in Fig. 5.

Table 1. Subbands in low-frequency range.

1st	band	$0 - P_1$
2nd	band	$P_1 - P_2$
...		
K_1-th	band	$P_{K_1-1} - P_{K_1}$

Next, the signal is filtered within the bandwidth of [930–8000] Hz. The filtered signal is shown in Fig. 6. As seen in Fig. 6, this signal exhibits the periodicity.

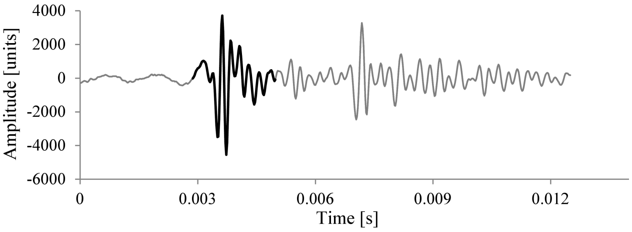

Fig. 6. The phoneme /p'/ signal with frequencies from the bandwidth of 930–8000 Hz.

The dark curve shown in Fig. 6 indicates the chosen period on the basis of which the synthesizer model will be created. The magnitude response of the selected period is calculated. The obtained magnitude response is divided into 10 frequency bands that are shown in Fig. 7.

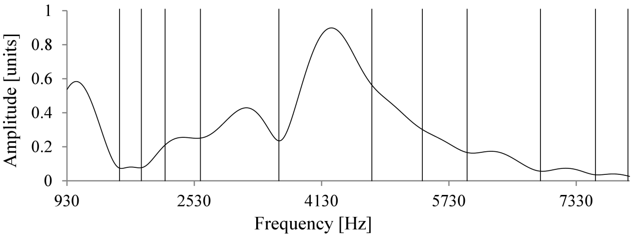

Fig. 7. The phoneme /p'/ signal spectrum with frequencies from the bandwidth of 930–8000 Hz.

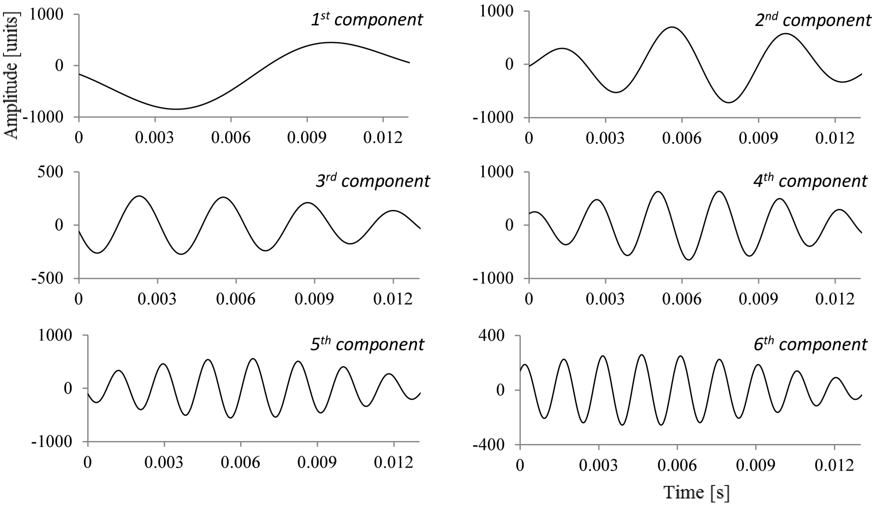

Fig. 5. The phoneme /p'/ components with frequencies from the bandwidth of 1–930 Hz.

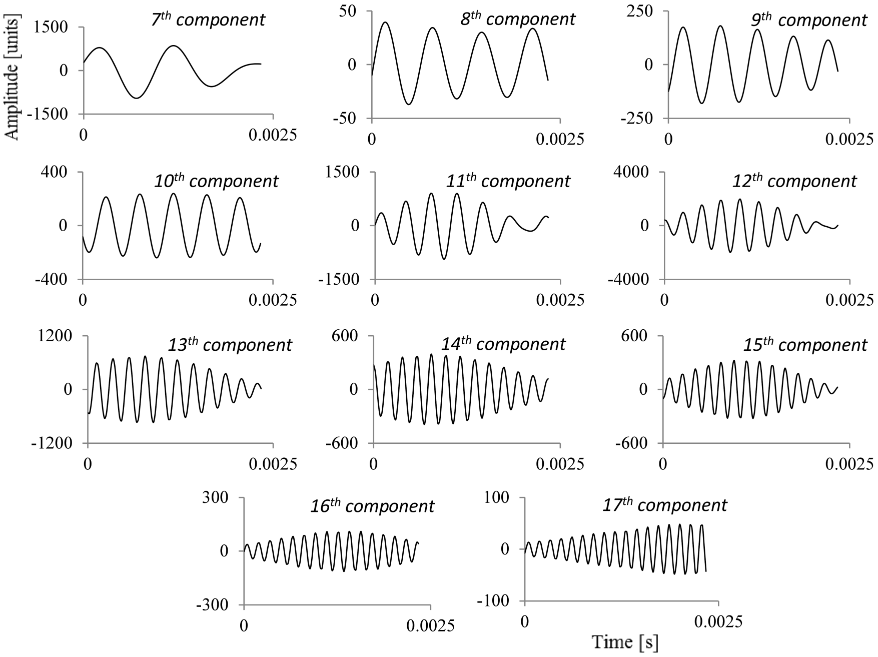

Fig. 8. The phoneme /p'/ components with frequencies from the bandwidth of 790–8000 Hz.

For each of the frequency bands, the inverse Fourier transform is applied. The obtained signals are presented in Fig. 8.

After dividing the signal into components in low and high frequency ranges, 17 signals of a simple form are obtained. The lengths of these signals are 0.013 s and 0.0024 s, respectively. Each of the obtained signals is modeled by formula (1). The parameters of the 5th–8th component impulse responses are shown in Table 2.

The inputs $\{u(n)\}$ of the MISO system are presented in Fig. 9.

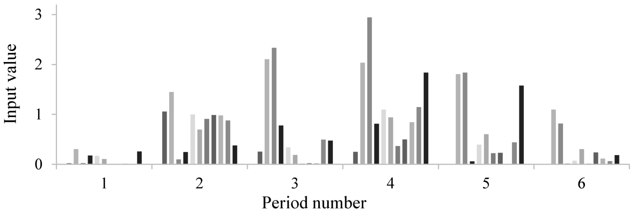

Fig. 9. The input values of the phoneme /p'/.

In order to evaluate the accuracy of modelling, the Fourier transforms of the real data and output of the MISO system have been compared. The magnitude response shows only small differences (see Fig. 9).

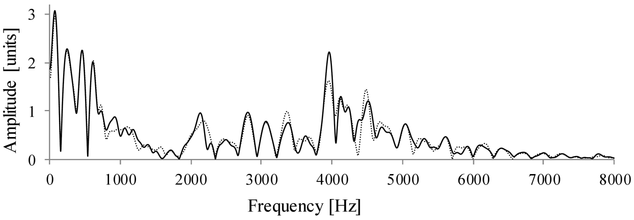

Fig. 10. The spectra of the true phoneme /p'/ and its model (solid line – the true speech signal spectrum, dotted line – the modeled signal spectrum).

The mean absolute error (MAE) is employed in the model evaluation (Chai, Draxler, 2014). The MAE is calculated by the following formula:

$$\mathrm{MAE} = 100\% \cdot \frac{1}{Q} \sum_{q=1}^{Q} \left| S_q - \widehat{S}_q \right|, \tag{4}$$

where S_q is the q-th value of the spectrum of the true phoneme, $\widehat{S}_q$ – the q-th value of the spectrum of the true phoneme.

Table 2. The parameters of the 5th–8th component impulse responses of the phoneme /p'/.

Component number	f	λ	a_1	a_2	a_3	a_4	φ_1	φ_2	φ_3	φ_4
5	605	−504	222.7	7.11	0.040	0.0005	2.89	−2.35	1.59	−2.06
6	729	−530	168.9	4.10	0.011	0.0002	0.85	0.76	−2.14	0.88
7	999	−3101	0	162.8	7.89	0.24	0	0.86	−1.55	1.29
8	1718	−3098	0	18.0	0.64	0.02	0	−0.84	2.31	−0.87

We carried out the modelling using 15 utterances for all voiceless stop consonants. In order to show how the method deals with noise, we add random noise to the consonant phoneme signals. The signal-to-noise ratio (SNR) of the noisy signal is equal to 20 dB. The MAE values of the estimated signal spectrum and its confidence intervals are presented in Table 3.

Table 3. The MAE for the estimated voiceless stop consonant phoneme signal spectrum.

Phoneme	Real-valued signal		Noisy signal	
	MAE	Confidence intervals	MAE	Confidence intervals
/k/	5.71 %	[5.17, 6.26]	6.17 %	[5.59, 6.75]
/k'/	6.98 %	[6.30, 7.65]	7.38 %	[6.74, 8.02]
/p/	5.96 %	[5.32, 6.61]	7.12 %	[5.95, 8.29]
/p'/	6.52 %	[5.94, 7.11]	7.36 %	[6.35, 8.37]
/t/	6.24 %	[5.46, 7.01]	6.78 %	[5.94, 7.62]
/t'/	6.67 %	[5.86, 7.48]	7.11 %	[6.37, 7.85]

The spectrum estimation errors (see Table 3) revealed that quality difference between the models of the real-valued signal and the noisy signal is small. The average MAE for the estimated signal spectrum of real-valued signals is equal to 6.35%, the average MAE for the estimated signal spectrum of noisy signals is equal to 6.99%. The small spectrum estimation errors revealed that quality of models is good.

Examples of the synthesized speech are uploaded on the website: http://audioakustyka.org/voiceless-stop-consonant-modeling-and-synthesis/.

5. Conclusions

Voiceless stop consonant phoneme modelling framework based on a phoneme modelling in low-frequency range and high-frequency range separately is proposed. A new limit point searching three-step algorithm and a new near points rejection algorithm is given in this paper.

The simulation has revealed that the proposed modelling framework is able to reconstruct the signal with noise. The average MAE of the estimated signal spectrum is equal to 6.35% for real-valued signals and 6.99% for noisy signals. The SNR of the noisy signal was equal to 20 dB.

Small estimation errors indicate that the phoneme model obtained by the proposed methodology is sufficiently good. High modelling quality was achieved due to:

1) the high order of quasi-polynomial order,
2) separate excitation impulse sequences for each component.

The study shows that it is possible to develop the consonant signal mathematical model that generates naturally sounding sound. In the future, such models of other consonant groups are to be developed.

Acknowledgments

Research partially sponsored by the Polish National Commission for UNESCO Scheme (fellowship grant financed by the Ministry of Science and Higher Education) and the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.

References

1. *AGH Corpora, Audiovisual Polish Speech Corpus*, http://www.dsp.agh.edu.pl/en:resources:korpusav#.wdgpivxrpln (accessed Jan., 2017).

2. Bergier M. (2014), *Instruction and production training practice on awareness raising, awareness in action: the role of consciousness in language acquisition*, [in:] *Second language learning and teaching*, Łyda A., Szczęśniak K. [Eds.], Springer International Publishing, doi: 10.1007/978-3-319-00461-7_7.

3. Birkholz P. (2013), *Modeling consonant-vowel coarticulation for articulatory speech synthesis*, PLoS ONE **8**, 4, e60603, doi: 10.1371/journal.pone.0060603.

4. Brocki Ł., Marasek K. (2015), *Deep belief neural networks and bidirectional long-short term memory hybrid for speech recognition*, Archives of Acoustics, **40**, 2, 191–195, doi: 10.1515/aoa-2015-0021.

5. Chai T., Draxler R.R. (2014), *Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature*, Geoscientific Model Developement, **7**, 1247–1250, doi: 10.5194/gmd-7-1247-2014.

6. Czyżewski A., Kostek B., Bratoszewski P., Kotus J., Szykulski M. (2017), *An audio-visual corpus for multimodal automatic speech recognition*, J. of Intelligent Information Systems, **1**, 1–26, doi: 10.1007/s10844-016-0438-z.

7. Demenko G., Mobius B., Klessa K. (2010), *Implementation of Polish speech synthesis for the boss system*, Bulletin of the Polish Academy of Sciences Technical Sciences, **58**, 3, doi: 10.2478/V10175-010-0035-1, http://bulletin.pan.pl/(58-3)371.pdf.

8. Domagała P., Richter L. (1994), *Automatic discrimination of Polish stop consonants based on bursts analysis*, Archives of Acoustics, **19**, 2, 147–159, http://acoustics.ippt.pan.pl/index.php/aa/article/-view/1084.

9. Driaunys K., Rudžionis V., Žvinys P. (2005), *Analysis of vocal phonemes and fricative consonant discrimination based on phonetic acoustics features*, Information Technology and Control, **34**, 3, 257–262.

10. Dziubiński M., Kostek B. (2005), *Octave error immune and instantaneous pitch detection algorithm*, Journal of New Music Reseach, **34**, 3, 273–292.

11. Gardzielewska H., Preis A. (2007), *The intelligibility of Polish speech synthesized with a new sinewave synthesis method*, Archives of Acoustics, **32**, 3, 579–589.

12. Gussmann E. (2007), *The phonology of Polish*, New York: Oxford University Press.

13. Igras M., Ziółko B., Jadczyk T. (2013), *Audiovisual database of Polish speech recordings*, Studia Informatica, **33**, 2b, 163–172.

14. Jadczyk T., Ziółko M. (2015), *Audio-visual speech processing system for Polish with dynamic Bayesian Network Models*, Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015), Barcelona, Spain, July 13–14, Paper No. 343.

15. Jassem W. (2003), *Polis*h, Journal of the International Phonetic Association, **33**, 103–107.

16. Johannessen J.B., Hagen K., Priestley J.J., Nygaard L. (2007), *An advanced speech corpus for Norwegian*, Proceedings of the 16th Nordic Conference of Computational Linguistics Nodalida-2007, 29–36, Tartu, Estonia, ISBN 978-9985-4-0513-0.

17. Korzinek D., Marasek K., Brocki Ł. (2011), *Automatic transcription of Polish radio and television broadcast audio*, Intelligent Tools for Building a Scientific Information Platform, Vol. 467, pp. 489–497, Springer.

18. Krynicki G. (2006), *Contrasting selected aspects of Polish and English phonetics*, http://ifa.amu.edu.pl/∼krynicki/my_pres/my_pres_6c.htm (accessed Jan. 2017).

19. Labarre T. (2011), *LING550: CLMS project on Polish*, http://www.academia.edu/5332895/ling550_clms_project_on_polish.

20. Ladefoged P., Disner S.F. (2012), *Vowels and consonants*, 3rd Ed., Ladefoged P. [Ed.], Wiley-Blackwell, Chichester.

21. Oliver D., Szklanny K. (2006), *Creation and analysis of a Polish speech database for use in unit selection synthesis*, http://syntezamowy.pjwstk.edu.pl/publikacje/lrec2006.pdf (accessed Jan. 2017).

22. Oostdijk N. (2000), *The spoken Dutch corpus. Overview and first evaluation*, Proceedings of LREC 2000, pp. 887–894, Athens, Greece.

23. Pinnis M., Auziňa I. (2010), *Latvian text-to-speech synthesizer*, Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010, pp. 69–72, Riga, Latvia: IOS Pres, doi:10.3233/978-1-60750-641-6-6.

24. Pinnis M., Auziňa I., Goba K. (2014), *Designing the Latvian speech recognition corpus*, Proceedings of 9th International Conference on Language Resources and Evaluation, LREC'14, pp. 1547–1553.

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

26. Pyž G., Šimonytė V., Slivinskas V. (2014), *Developing models of Lithuanian speech vowels and semivowels*, Informatica, **25**, 1, 55–72.

27. Raitio T., Lu H., Kane J., Suni A., Vainio M., King S., Alku P. (2014), *Voice source modelling using deep neural networks for statistical parametric speech synthesis*, [in:] *European Signal Processing Conference*, 6952838, European Signal Processing Conference, EUSIPCO, pp. 2290–2294, 22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, United Kingdom, 1–5 September.

28. Raškinis A., Dereškeviciutė S. (2007), *An analysis of spectral attributes, characterizing the interaction of lithuanian voiceless velar stop consonants with their pre- and postvocalic context*, Information Technology and Control, **36**, 1, 68–75.

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

30. SAMPA Homepage (2005) [in Polish], http://www.phon.ucl.ac.uk/home/sampa/polish.htm (last revised 2005; accessed Jan. 2017).

31. SAMPA Homepage (2005), http://www.phon.ucl.ac.uk/home/sampa/index.html (last revised 2005; accessed Jan. 2017).

32. Sasirekha D., Chandra E. (2012), *Text to speech: a simple tutorial*, International Journal of Soft Computing and Engineering (IJSCE), **2**, 1, 275–278.

33. Stănescu M., Cucu H., Buzo A., Burileanu C. (2012), *ASR for low-resourced languages: building a phonetically balanced Romanian speech corpus*, Proceedings of 20th European Signal Processing Conference, pp. 2060–2064.

34. Stevens K.N. (1993), *Modelling affricate consonants*, Speech Communication, **13**, 1–2, 33–43.

35. Tabet Y., Boughazi M. (2011), *Speech synthesis techniques. A survey*, 7th International Workshop on Systems, Signal Processing and Their Applications (WOSSPA), pp. 67–70.

36. Tamulevičius G., Kaukėnas J. (2016), *Adequacy analysis of autoregressive model for Lithuanian semivowels*, Advances in Information, Electronic and Electrical Engineering (AIEEE), 2016 IEEE 4th Workshop on, doi: 10.1109/AIEEE.2016.7821825.

37. Tokuda K., Nankaku Y., Toda T., Zen H., Yamagishi J., Oura K. (2013), *Speech synthesis based on hidden Markov Model*, Proceedings of the IEEE, **101**, 5, 1234–1252.

38. Upadhyaya P., Farooq O., Abidi M.R., Varshney P. (2015), *Comparative study of visual feature for bimodal Hindi speech recognition*, Archives of Acoustics, **40**, 4, 609–619, doi: 10.1515/aoa-2015-0061.

39. VoxForge (2017), http://www.voxforge.org/home/downloads (accessed Jan. 2017).

40. Żelasko P., Ziółko B., Jadczyk T., Skurzok D. (2016), *AGH corpus of Polish speech*, Language Resources and Evaluation, **50**, 3, 585–601, doi: 10.1007/S10579-015-9302-Y.

41. Zen H., Tokuda K., Black A.W. (2009), *Statistical parametric speech synthesis*, Speech Communication, **51**, 11, 1039–1064.

42. Ziółko B., Gałka J., Suresh M., Wilson R., Ziółko M. (2009), *Triphone statistics for Polish language*, Human Language Technology: Challenges of the Information Society, LTC 2007, Lecture Notes in Computer Science, Vol. 5603, pp. 63–73, Springer, Berlin, Heidelberg.

43. Ziółko B., Ziółko M. (2011), *Time durations of phonemes in Polish language for speech and speaker recognition*, Human Language Technology. Challenges for Computer Science and Linguistics. Lecture Notes in Computer Science, Vol. 6562, 105–114, Springer Verlag.