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

Grażina KORVEL(1), Bożena 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; IGRAŞ et al., 2013; JADCYK, ZIÓŁKO, 2015; JOHANNESSEN et al., 2007; KORZINEK et al., 2011; OOST-DIJK, 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 Po-
lish, 2005). As pointed out by the authors of that pa-
per, 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 signifi-
cant reduction of information carried by speech signal
(GARDZIELEWSKA, PREIS, 2007).

A voice source modelling method based on predict-
ing the time domain glottal flow waveform using
a DNN (Deep Neural Network) is described in very re-
cent sources (RAITIO et al., 2014). TAMULEVIČIUS and
KAUKÉNAS (2016) apply Autoregressive model param-
eter estimation technique for modelling of semivowels.
Contrarily, much less attention has been paid to for-
mant speech synthesis. The main reason is that the
synthesized speech quality does not achieve the natu-
real speech quality yet (SASIREKHA, CHANDRA, 2012;
Tabet, Boughazi, 2011). Formant synthesizers have advan-
tages against the concatenative ones. The speech pro-
duced by them can sufficiently be intelligible even at
high speed (Tabet, Boughazi, 2011). They can con-
tral 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 bet-
ter 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 syn-
thesis. The signals of consonant phonemes are more
difficult than those of vowels and semivowels. For ex-
ample, no previous study has considered Lithuanian
stop consonant /k/ of Lithuanian. The phonology of
Polish was described in many sources (e.g. 1994;
KRINICKI, 2006) some contrasting aspects of Po-
lish 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 scenar-
ios, including phonetically balanced 4.5 h subcorpus
recorded in an anechoic chamber, which may be use-
ful for extracting material for carrying out evaluation
tests (AGH Corpora; ŽELASKO et al., 2016). The pho-
netical statistics were collected from several Polish cor-
pora (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 recog-
nition and speaker identification systems. This feature
could be used in phoneme parametrization and mod-
delling (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 or-
gans producing sounds of the given language are con-
sidered in these papers. In the current research sound
is described in terms of acoustical properties, i.e. sig-
nal 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 conso-
nant 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 (DOMAGALA, 1994;
KRINICKI, 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 syn-
thesis 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 mod-
els can also be adapted to other similar problems, for
example treating language disorders, speech recogni-
tion, 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 ex-
periments. Conclusions are presented in the last sec-

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.

\[
h(t) = e^{\lambda t}(a_1 \sin(2\pi ft + \varphi_1) + a_2 t \sin(2\pi ft + \varphi_2) + a_3 t^2 \sin(2\pi ft + \varphi_3) + a_4 t^3 \sin(2\pi ft + \varphi_4)),
\]

where \( t \in R^+ \cup \{0\}, \lambda > 0 \) – the damping factor, \( f \) – the frequency, \( a_k \) – amplitude, \( \varphi_k (-\pi \leq \varphi_k < \pi) \) – 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 \( \text{cross point} \).

3) From the \( \text{cross point} \), we will go down until we reach a minimum. Such a point will be a \( \text{limit point} \).

Figure 2 shows the graphical representation of limit point searching. Note that the frequency spectrum from the range \([0, 2000] \text{ Hz}\) 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 \([1, f_{\text{lim}}] \) 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 Pyż 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:

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

(2)

where $\text{FT}(m)$ – 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:

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

(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.

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

The signal is filtered with a filter with the bandwidth of $[f_{\text{limit}}, 8000]$ 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 Pyż 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>
<thead>
<tr>
<th>Band</th>
<th>Frequency Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st band</td>
<td>$0 - P_1$</td>
</tr>
<tr>
<td>2nd band</td>
<td>$P_1 - P_2$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$K_1$-th</td>
<td>$P_{K_1} - P_{K_1}$</td>
</tr>
</tbody>
</table>

Table 1. Subbands in low-frequency range.

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.
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.

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).

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

\[
\text{MAE} = 100\% \cdot \frac{1}{Q} \sum_{q=1}^{Q} |S_q - \hat{S}_q|, \tag{4}
\]

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 true phoneme.

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

<table>
<thead>
<tr>
<th>Component number</th>
<th>( f )</th>
<th>( \lambda )</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( a_3 )</th>
<th>( a_4 )</th>
<th>( \varphi_1 )</th>
<th>( \varphi_2 )</th>
<th>( \varphi_3 )</th>
<th>( \varphi_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>605</td>
<td>−504</td>
<td>222.7</td>
<td>0.040</td>
<td>0.0005</td>
<td>2.89</td>
<td>−2.35</td>
<td>1.59</td>
<td>−2.06</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>729</td>
<td>−530</td>
<td>168.9</td>
<td>4.10</td>
<td>0.011</td>
<td>0.0002</td>
<td>0.85</td>
<td>0.76</td>
<td>−2.14</td>
<td>0.88</td>
</tr>
<tr>
<td>7</td>
<td>999</td>
<td>−3101</td>
<td>0</td>
<td>162.8</td>
<td>7.89</td>
<td>0.24</td>
<td>0</td>
<td>0.86</td>
<td>−1.55</td>
<td>1.29</td>
</tr>
<tr>
<td>8</td>
<td>1718</td>
<td>−3098</td>
<td>0</td>
<td>18.0</td>
<td>0.64</td>
<td>0.02</td>
<td>0</td>
<td>−0.84</td>
<td>2.31</td>
<td>−0.87</td>
</tr>
</tbody>
</table>

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

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

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).
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.

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Real-valued signal</th>
<th>Noisy signal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>Confidence</td>
</tr>
<tr>
<td>/k/</td>
<td>5.71 %</td>
<td>[5.17, 6.26]</td>
</tr>
<tr>
<td>/k'/</td>
<td>6.98 %</td>
<td>[6.30, 7.65]</td>
</tr>
<tr>
<td>/p/</td>
<td>5.96 %</td>
<td>[5.32, 6.61]</td>
</tr>
<tr>
<td>/p'/</td>
<td>6.52 %</td>
<td>[5.94, 7.11]</td>
</tr>
<tr>
<td>/t/</td>
<td>6.24 %</td>
<td>[5.46, 7.01]</td>
</tr>
<tr>
<td>/t'/</td>
<td>6.67 %</td>
<td>[5.86, 7.48]</td>
</tr>
</tbody>
</table>

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


