The tendencies and features of an excitation signal modeling in digital devices of the speech analysis and synthesis

Alexander Rybolovlev, Sergej Zabirnik (1), Mihail Galkin (2)

(1) Academy of Special communication of Russia, 35, Priborostroitelnaya str., Oryol, Russia  
rybolovlev@rambler.ru

(2) Group of government communication, Syzran, Russia  
galkin@front.ru

Abstract

The modern level of development of digital devices of the analysis and synthesis of speech based on a method of a linear prediction, is characterized by the large variety of ways of allocation, approximation and digital representation of coded parameters. The practice of speech coding shows, that more than 70 % of information resources selected on coding of the current speech frame, is used for the description of an excitation signal of the synthesizing digital filter.

The definition of laws and tendencies of development of the specified procedures will facilitate to the developers of perspective speech codecs the search of the algorithms ensuring increase of an optimality of analog-digital transformation of a speech signal in relation to elected criterion functions. In offered paper are considered the tendencies and features of an excitation signal modeling in digital devices of the speech analysis and synthesis.

Introduction

Feature of speech transfer in digital communication systems with using of source coding methods is transfer on a communication channel not the signal, but model of its formation. When coding with a linear prediction (LPC) is used, the model includes parameters of some filter equivalent to a speech path, and parameters of an excitation signal for this filter. The task of coding on the transmitting site of the communication line consists in an estimation of filter and excitation signal parameters, while a task of decoding on the receiving site consists in a gating of excitation signal through the filter on which output the reconstructed speech is formed. Various variants of coding algorithms differ one from another a set of filter parameters, a method of excitation signal formation and other details. Figure 1 explains the essence of such approach.

![Diagram](image)

Figure 1 The analysis and synthesis of speech by means of a linear prediction method

In case of LP-coding the forming of an excitation signal is made by gating an initial speech signal through the filter - analyzer (before it the speech signal is digitized and segmented). The following model of the analysis is used:

\[ E(z) = S(z)A(z), \]

where \( A(z) \) – transfer function of the analysis filter

\[ A(z) = 1 - \sum_{i=1}^{M} a(i)z^{-i}, \]

\( S(z) \) – initial speech signal, \( M \) – prediction order.

Thus, the deferred signal \( e(n) \) is formed on an filter output

\[ e(n) = s(n) - \sum_{i=1}^{M} a(i)s(n-i) \]

This signal represents an error of prediction, and is the elementary model of an excitation signal. Signal
\( e(n) \) is quantized, coded and passed in a communication channel.

Redundancy of a speech signal at passage through the filter - analyzer is appreciably decreased. Internal correlation in a signal are eliminated. Signal has a spectrum with flat spectral envelope. Sometimes this process is called “spectrum whitewash” or “inverse filtering”. Because the filter has the frequency characteristic, reverse to a signal spectrum, it names the “inverse filter”.

Graphic representation of the given algorithm is resulted in figure 2.

Input signal Analysis filter Residual signal
Spectral representation

Time representation

\[
S(z) \rightarrow A(z) \rightarrow E(z)
\]

Figure 2. Short-term redundancy elimination by transfer of a speech signal (vowel sound)

Procedure of speech decoding consists in a gating the accepted excitation signal of \( e(n) \) through synthesizing filter which parameters are transferred together with a excitation signal. The synthesizing filter has the same structure, as analyzing, and is defined by the same set of parameters (prediction coefficient \( a(i) \), reflection coefficient \( k(i) \), log-area ratio \( r(i) \) or line spectrum pair (LSP).

The model of synthesis is described by the following mathematical expression:

\[
S(z) = E(z) \frac{1}{A(z)}.
\]

The algorithms of speech coding described above appreciably differ from the algorithms of a digital speech compression realized practically. The given scheme of a linear prediction – a short-term prediction (STP) does not provide a sufficient degree of speech redundancy elimination since the residual signal still contains quasi-periodical components, first of all - the pitch.

Therefore, in addition to a short-term prediction, the long-term prediction (LTP) is used. LTP appreciably eliminates residual redundancy and approaches the rest of a prediction under the statistical characteristics to white noise.

In view of that the pitch is characterized by only two parameters: amplitude and the period, transfer function of long-term prediction filter looks as follows:

\[
P(z) = 1 - Gz^{-d},
\]

where \( G \) – the gain factor describing amplitude of the pitch pulses, \( d \) – the delay determining the pitch period.

If the remaining signal of a short-term prediction \( e(n) \) moves on an input of the long-term prediction filter on its output the long-term prediction residual \( f(n) \) turns out:

\[
f(n) = e(n) - G \cdot e(n - d).
\]

The signal \( f(n) \) is close to white Gaussian noise that facilitates economic formation of excitation signal parameters. Parameters of a long-term prediction \( G \) and \( d \) can be determined, from a condition of minimization mean square value of a prediction error \( f(n) \) on some interval making 20...25 % from duration of the transmitted segment of speech. The delay \( d \) usually consists within the limits of 20…160 sampling interval that corresponds to a range of pitch frequencies of 50…400 Hz.

Use of the rest of a prediction \( f(n) \) as a signal of excitation appears insufficiently effective as demands for coding too big number of bits. Therefore more economic methods (on loading a communication channel, but not on computing expenses) of excitation signal formation find practical application.

Modern methods of a speech compression assume multistage analysis of an input speech signal with the purpose of formation of its model, the optimal on loading a communication channel, at preservation of required quality of synthetic speech (figure 3).

![Figure 3. Multistage formation of a speech signal model](image-url)
If procedures of the short-term and long-term analysis are similar in all modern speech coders, then methods of approximation of the second residual signal \( f(n) \) in various applications can differ essentially.

The signal of excitation approximating the LTP residual, is modeled as the certain number of pulses on an interval of the excitation frame, 20...50 \% making usually from duration of a transmitted segment of speech.

In a method of multipulse excitation (MPE) it is optimized both position, and amplitudes of pulses. The number of pulses can be chosen small (3-5), that there is less than number of signal samples on an excitation frame.

The signal of excitation in the multipulse approach is defined as:

\[
e(n) = \sum_{i=0}^{M-1} g(i) \delta(n - m(i)), \quad n = 0, ..., N - 1, \tag{7}
\]

where \( g(i) \) – amplitudes of pulses, \( m(i) \) – positions of pulses, \( N \) – number of the pulses modeling excitation sequence.

The basic advantage of multipulse excitation is that it is defined for any speech segment. Thus it is not required knowledge neither of character of a speech segment (voiced, unvoiced), nor of the pitch period.

In a method of excitation by regular sequence of pulses – Regular-Pulse Excitation (RPE), the positional relationship of pulses is predetermined beforehand - the grid of equidistant pulses is used, and the arrangement of this grid is optimized within the limits of the excitation frame and amplitudes of pulses. Method RPE is less effective in comparison with MPE, however the algorithm of processing at RPE is much easier.

In a method of stochastic coding, or a method of Code-Excited Linear Prediction (CELP), with a version of excitation by the vector sum (Vector Sum Excited Linear Prediction - VSELP), the most suitable vector of excitation is selected from predetermined code book, or the code dictionary containing usually 2N, \( N = 7...10 \), quasirandom vectors of the set length with the elements normalized on amplitude. The amplitude of a vector of excitation is coded separately according to loudness of a transmitted element of speech.

The effective method of excitation by sequence of binary pulses with transformation (Transformed Binary Pulse Excitation – TBPE) in which a signal of excitation is the sequence equidistant on time and quasirandom on a sign (with amplitudes \( \pm 1 \)) pulses multiplied on some matrix of transformation.

**The features of an excitation signal modeling in popular speech codecs**

Speech codec LPC-10

The given algorithm of speech production is accepted as the federal standard of speech transfer in USA: “analog-digital conversion of speech on the basis of a linear prediction with a speed of 2400 bits per second”. On the transmitting side the analog speech signal is filtered by the band-pass filter and is exposed to analog-digital conversion. Then there is a predistortion of a speech signal that reduces requirements to integrated accuracy at an estimation of prediction parameters. The linear prediction analyzer defines reflection coefficient and the characteristic of signal energy on an interval of the analysis (130 samples, 22.5 ms). If a segment is voiced the pitch period is defined; if unvoiced, whether that is defined, is the frame unvoiced or it is a pause of speech. Then the root-mean-square amplitude, reflection coefficient, pitch period and characteristic of segment type are coded. In view of antinoise coding and the bit of synchronization 54-bits frame is formed. The decoder forms a stream with a bit rate of 2400 bits per second.

Formation of an excitation signal occurs on the reception side of the communication line. The decoded parameters are transferred in the block of interpolation where they smooth out and interpolated. The interpolated values are used at synthesis of segments of a speech signal. Speech is synthesized by the filter of 10-th order on which input the sequence of pulses of the excitation (Table 1) repeated with the pitch period moves.

<table>
<thead>
<tr>
<th>index number</th>
<th>pulse amplitude</th>
<th>index number</th>
<th>pulse amplitude</th>
<th>index number</th>
<th>pulse amplitude</th>
<th>index number</th>
<th>pulse amplitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>249</td>
<td>11</td>
<td>-82</td>
<td>21</td>
<td>-82</td>
<td>31</td>
<td>-29</td>
</tr>
<tr>
<td>2</td>
<td>-262</td>
<td>12</td>
<td>376</td>
<td>22</td>
<td>-123</td>
<td>32</td>
<td>-21</td>
</tr>
<tr>
<td>3</td>
<td>363</td>
<td>13</td>
<td>288</td>
<td>23</td>
<td>-39</td>
<td>33</td>
<td>-18</td>
</tr>
<tr>
<td>4</td>
<td>-362</td>
<td>14</td>
<td>-65</td>
<td>24</td>
<td>65</td>
<td>34</td>
<td>-27</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>15</td>
<td>-20</td>
<td>25</td>
<td>64</td>
<td>35</td>
<td>-31</td>
</tr>
<tr>
<td>6</td>
<td>367</td>
<td>16</td>
<td>138</td>
<td>26</td>
<td>19</td>
<td>36</td>
<td>-22</td>
</tr>
<tr>
<td>7</td>
<td>79</td>
<td>17</td>
<td>-62</td>
<td>27</td>
<td>16</td>
<td>37</td>
<td>-12</td>
</tr>
<tr>
<td>8</td>
<td>78</td>
<td>18</td>
<td>-315</td>
<td>28</td>
<td>32</td>
<td>38</td>
<td>-10</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>19</td>
<td>-247</td>
<td>29</td>
<td>19</td>
<td>39</td>
<td>-10</td>
</tr>
<tr>
<td>10</td>
<td>-277</td>
<td>20</td>
<td>-78</td>
<td>30</td>
<td>-15</td>
<td>40</td>
<td>-4</td>
</tr>
</tbody>
</table>
In the given table are represented 40 relative amplitudes of excitation pulses determined by the standard for coder LPC-10. If the pitch period is equal 40, all 40 values are used during the pitch period. If the pitch period is more 40 the first 40 values of excitation pulses choose from the table, and the others are equated to zero. If the period of the basic tone is less 40 (between 20 and 39) superfluous values are rejected. The given transformation is carried out for each interpolated pitch period on a synthesized segment. At transition from one voiced segment on the following pulses of excitation which were not placed on a segment are remembered and located in the beginning of an excitation signal of the following segment. After that the periodic excitation signal for the following segment is formed.

Unvoiced sounds are synthesized by the filter of 4-th order with pseudorandom excitation (thus the choice of random numbers from same table with an any index point and an any increment is realized).

Pauses are synthesized by the filter excited by periodic sequence of pulses (classification of pauses, as voiced segments) and pseudorandom sequence of pulses (classification of pauses, as unvoiced segments).

The excitation signal is transformed by the phase filter raising quality of speech sounding. The excitation signal has a constant level, therefore an output signal of a synthesizer is scaled in the amplifier so that to correspond to root-mean-square value of amplitude.

After amplification the signal is exposed to a post filtration and digital-to-analog conversion.

Speech codec LPC-LTP-MPE

Algorithm LPC-LTP-MPE (coding of speech on the basis of a linear prediction with long-term multipulse excitation) is used in satellite communication system INMARSAT AERONAUTICAL.

The coder provides speed 9.6 kbit/s. Thus the low delay of transfer and high quality of speech is provided.

On the transmitting side the coder's analyzer allocates the following parameters:

- 10 linear prediction coefficients \((a_{10})\) and 10 reflection coefficients \((k_{10})\).
- Two long-term prediction coefficient: a delay \(d\) and gain factor \(G\).
- 30 parameters of a multipulse excitation signal: 15 positions of excitation pulses \((m_{1}...m_{15})\) and 15 values of their amplitudes \((g_{1}...g_{15})\).

The signal of multipulse excitation is defined for each of 5 subframes (32 samples, 4 ms). It represents sequence of three pulses with non-uniformly distributed intervals and various amplitudes. Amplitudes of pulses and their position in time are defined by a method "analysis by synthesis".

With use of found LPC and LTP coefficients synthesis of a signal on each of subframe is carried out:

\[
\hat{s}(n) = G a(n - d) + \sum_{i=1}^{15} a_i \hat{s}(n - i). \tag{8}
\]

Values of signals samples \(a(n - d)\) and \(\hat{s}(n - i)\) for which the argument is less than zero, undertake from previous subframes.

Further the signal of a prediction error is calculated:

\[
e(n) = s(n) - \hat{s}(n), \quad 0 \leq n \leq 31 \tag{9}
\]

The pulse characteristic \(h(n)\) of the synthesizing filter is calculated. Besides full energy of the pulse characteristic \(E(n)\) which is used for normalization of cross-correlation function is defined, at the following entry conditions: \(h(0) = 0, E(31) = 1\).

\[
h(n) = \sum_{i=1}^{31} a_i h(n - i), \quad 1 \leq n \leq 31; \tag{10}
\]

Cross-correlation function of a prediction error and the pulse characteristic is calculated:

\[
R_{w}(n) = \sum_{i=1}^{31} e(i)h(i - n), \quad 1 \leq n \leq 31; \tag{11}
\]

Position of an excitation pulse \(m(i)\) on a maximum of the relation \((R_{w}(n))^{2}/E(n)\) is defined:

\[
m_i \in \{0, 31\}, \quad \left( R_{w}(m) \right)^{2}/E(n) \geq \left( R_{w}(n) \right)^{2}/E(n), \quad \forall n = 0, ..., 31 \tag{12}
\]

The amplitude of an excitation pulse is defined under the formula 13:

\[
g_i = R_{w}(m_i)/E(m_i) \tag{13}
\]

Coding of parameters of speech production model is carried out according to tables of coding and decoding. Exception makes coding amplitudes of excitation signal pulses for subframes which is carried out as follows:

- Weighing (alignment) of pulses amplitudes.
- Division of the weighed value of excitation pulse amplitude into the regulating coefficient which is taking into account influence of one subframe on the following.
- Coding the value of excitation pulse amplitude according to the table of coding.
- Updating the regulating coefficient.

The found values of the coded parameters of a speech signal are united in the frames. Distribution of bits within the limits of the frames is resulted in table 2.

Decoding of excitation pulses amplitudes and other parameters, and also synthesis of speech on them is made on the reception side in return sequence.
Table 2: Bits distribution in the frame of speech codec LPC-LTP-MPE

<table>
<thead>
<tr>
<th>Transmitted parameters</th>
<th>Amount of bits</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters of the short-term prediction filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial correlation coefficients</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Parameters of the long-term prediction filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay, ( d )</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Gain, ( G )</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Parameters of an excitation signal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position of excitation pulses (for each of five segments)</td>
<td>75</td>
<td>( m_1...m_3 ) — on ( 6 ) bit</td>
</tr>
<tr>
<td>Amplitudes of excitation pulses (for each of five segments)</td>
<td>55</td>
<td>( g_1...g_2 ) — on ( 4 ) bits; ( g_3 ) — 3 bits</td>
</tr>
<tr>
<td>Hamming code</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>In total on a segment (20 ms)</td>
<td>192</td>
<td></td>
</tr>
</tbody>
</table>

Standard DAMPS

In D-AMPS standard the VSELP method of coding is used.

The speech signal passes front-end processing: a filtration and segmentation. One segment makes 160 samples (20 milliseconds).

Parameters of an excitation signal are defined for each of subframes on 40 samples. The coding scheme includes filters - synthesizers of a short-term and long-term prediction and two code books and realizes a method “analysis by synthesis”. Each of code books of an excitation signal contains 128 code vectors, on 40 elements in everyone. All code vectors of one book are elements 7-dimensional linear subspace of 40-dimensional space formed by 7 basic vectors. And linear combinations coefficients specifying code vectors through vectors of basis have values \( \pm 1 \) or 1. Thus, each code book containing 128 vectors is set by 7 basic vectors and 128 code words (7-element vectors of linear combinations coefficients) with one-bit elements.

The excitation signal of a short-term prediction filter - synthesizer, is the sum of excitation vectors from two code books and a vector from an output of a long-term prediction filter - synthesizer (therefore the name of a method — “vector sum excitation”). Vectors of excitation from code books are multiplied on corresponding gain factors \( \gamma \) and \( \gamma_2 \). The input signal of a long-term prediction filter - synthesizer corresponds, depending on a type of a segment, to an output signal of the same filter or a total excitation signal of a short-term prediction filter - synthesizer. Parameters of an excitation signal (number of excitation vectors \( I_1 \) and \( I_2 \) from the first and second code books and corresponding gain factors \( \gamma_1 \) and \( \gamma_2 \)) are defined by criterion of a minimum root-mean-square mistakes on an output of a short-term prediction filter - synthesizer included in the coder.

Distribution of speech coder DAMPS output information is resulted in table 3.

Table 3: Bits distribution in the frame of speech codec DAMPS

<table>
<thead>
<tr>
<th>Transmitted parameters</th>
<th>Amount of bits</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters of the short-term prediction filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial correlation coefficients</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Amplitude factor (Energy of a segment) ( \rho )</td>
<td>5</td>
<td>Один на сегмент</td>
</tr>
<tr>
<td>Parameters of the long-term prediction filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay, ( d ) (for each of four subframes)</td>
<td>28</td>
<td>7 bits on each subframes</td>
</tr>
<tr>
<td>Excitation signal parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of excitation vectors ( I_1 ), ( I_2 ) from two code books (for each of four subframes)</td>
<td>56</td>
<td>( I_1, I_2 ) on 7 bits</td>
</tr>
<tr>
<td>Gain factors ( g_1 ), ( g_2 ) (for each of four subframes)</td>
<td>8</td>
<td>8 bits on each subframes (Some functions from ( g_1 ), ( g_2 ) are exposed to vector quantization)</td>
</tr>
<tr>
<td>In total on a segment (20 ms)</td>
<td>159</td>
<td></td>
</tr>
</tbody>
</table>

On the reception side the decoder carries out return operations. Under numbers \( I_1 \), \( I_2 \) from code books excitation vectors get out. They are multiplied on factors \( \gamma_1 \), \( \gamma_2 \) and are added with an output vector of a long-term prediction filter - synthesizer determined in parameters \( g \), \( d \). Further the excitation signal is filtered by a short-term prediction filter-synthesizer in the form of transversal filter. Parameters of the filter transform - from partial correlation coefficients \( k(i) \) pass to coefficients of a prediction \( a(i) \). For improvement of subjective quality of the synthesized speech the output signal of a filter - synthesizer is exposed to a digital adaptive post-filtration. On an output of the post-filter the reconstructed speech turns out.

Standard GSM

In standard GSM RPE-LTP method (Regular Pulse Excited Long Term Predictor) is used. The en-
trance signal is deformed, segmented (segments on 160 samples, 20 milliseconds), and is weighed by Hamming window. The estimation of an excitation signal is made on the smoothed rest of a prediction \( f_0 \) (signal from an output of the long-term prediction filter). Parameters of a excitation signal are estimated separately for every subframe on 40 samples.

The excitation signal of one subframes consist of 13 pulses following at regular intervals (three times big than a sampling interval of an initial signal), and having various amplitudes. For formation of an excitation signal a segment of the smoothed rest \( f(n) \) from 40 pulses are processed as follows. Last (fortieth) pulse is rejected, and the first 39 pulses are broken into three sequences: in the first - pulses 1, 4... 37, in the second - pulses 2, 5... 38, in the third - pulses 3, 6... 39. As an excitation signal that gets out of sequences which energy is more. Amplitudes of pulses are normalized in relation to a pulse with the greatest amplitude, and the normalized amplitudes are coded by three bits everyone at a linear scale of quantization. Absolute value of the greatest amplitude is coded by six bits in logarithmic scale. Position of an initial pulse of 13-element sequence is coded by two bits.

Distribution of the output coder information is resulted in table 4.

Table 4: Bits distribution in the frame of speech codec GSM

<table>
<thead>
<tr>
<th>Transmitted parameters</th>
<th>Amount of bits</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters of the short-term prediction filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-area ratio ( r_1...r_5 )</td>
<td>36</td>
<td>( r_1, r_2 ) - 06 bits; ( r_3, r_4 ) - on 5 bits; ( r_5 ) - on 4 bits; ( r_7, r_8 ) - on 3 bits</td>
</tr>
<tr>
<td>Parameters of the long-term prediction filter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay ( d ), gain factor ( g ) (for each of four subframes)</td>
<td>36</td>
<td>( g ) - 2 bits, ( d ) - 7 bits</td>
</tr>
<tr>
<td>Excitation signal parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of sequence ( n ), peak amplitude ( v ), normalized pulses amplitudes ( b_1...b_{13} ) (for each of four subframes)</td>
<td>188</td>
<td>( n ) - 2 bits, ( v ) - 6 bits, ( b_1 ) - 3 bits</td>
</tr>
<tr>
<td>In total on a segment (20 ms)</td>
<td>260</td>
<td></td>
</tr>
</tbody>
</table>

At decoding there are following operations. The block of excitation signal formation, using the accepted excitation signal parameters, restores 13-pulse excitation sequence for each of subframes of speech, including amplitudes of pulses and their arrangement in time. The generated excitation signal is filtered by a long-term prediction filter-synthesizer on which output the restored rest of a prediction of the STP filter-analyzer turns out. Last is filtered by a trellised STP filter-synthesizer, and parameters of the filter will preliminary be transformed from log-area ratio \( r(i) \), to a partial correlation coefficient \( k(i) \). The output signal of a STP filter - synthesizer is filtered (in the block of a post-filtration) by the digital filter restoring peak parities of frequency components of a speech signal. The signal on the postfilter output is a reconstructed speech.

CS-ACELP codec (recommendation ITU-T G.729)

The standard provides coding speech for the speed of 8 kbits per second. The coder uses algorithm of a linear prediction with excitation from an algebraic code and the conjugate structure of the quantizer.

The coder works with frames on 10 milliseconds (80 samples). Each frame is analyzed for allocation of CELP-model parameters (linear prediction filter's coefficients, indexes adaptive and fixed code books, gain factors). Distribution of bits at coding the specified parameters is resulted in table 5.

Table 5: Bits distribution in the frame of speech codec CS-ACELP

<table>
<thead>
<tr>
<th>Transmitted parameters</th>
<th>Amount of bits</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line spectrum pair</td>
<td>18</td>
<td>Two-cascade vector quantization with an interframe linear prediction</td>
</tr>
<tr>
<td>Delay of the adaptive code book</td>
<td>13</td>
<td>8 bits on the first subframe, 5 bits on the second subframe</td>
</tr>
<tr>
<td>Check of a delay</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Index of the fixed code book</td>
<td>26</td>
<td>13 bits on the first subframe, 13 bits on the second subframe</td>
</tr>
<tr>
<td>Record of the fixed code book</td>
<td>8</td>
<td>4 bits on the first subframe, 4 bits on the second subframe</td>
</tr>
<tr>
<td>Gain factor of the code book (a stage 1)</td>
<td>6</td>
<td>3 bits on the first subframe, 3 bits on the second subframe</td>
</tr>
<tr>
<td>Gain factor of the code book (a stage 2)</td>
<td>8</td>
<td>4 bits on the first subframe, 4 bits on the second subframe</td>
</tr>
<tr>
<td>In total on a segment (10 ms)</td>
<td>80</td>
<td></td>
</tr>
</tbody>
</table>

The signal of excitation is formed on the basis on “analysis by synthesis” method and represents set of adaptive and fixed code books vectors with gain factors corresponding to them. The choice of an excitation signal from the code book will consist of a choice of an optimum index of excitation and such gain factor concerning to it that the synthesized speech to the greatest degree corresponded to initial (original) speech and would result in minimization of the weighed rest of a prediction. Purpose of the adaptive
code book will consist in elimination of periodicity in a speech signal.

CELP-codec on 4.8 kbit/s (standard FS-1016)

The only standard is accepted as federal in the USA. The coder will consist of adaptive and stochastic code books, short-term and long-term prediction analyzers, the weighing filter, the block of minimization of the weighed prediction error and the coding device.

The signal is analyzed on the frame on 30 ms (everyone frame is divided on 4 subframes on 7.5 ms).

The excitation signal includes the sum of a adaptive code book vector $C_i$ and gain factor $g_i$ corresponding to it and a vector of stochastic code book $C_s$ with gain factor $g_s$. Distribution of bits in limits of subframes is resulted in table 6.

Table 6: Bits distribution in the frame of speech codec CELP

<table>
<thead>
<tr>
<th>Transmitted parameters</th>
<th>Amount of bits</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line spectrum pair</td>
<td>34</td>
<td>10 LSP on 3,4,4,4,4,3,3,3 bits accordingly</td>
</tr>
<tr>
<td>Vector of adaptive code book $C_i$</td>
<td>28</td>
<td>8 bits on the first sub-frame, 6 bits on the second sub-frame, 8 bits on the third sub-frame, 6 bits on the fourth sub-frame</td>
</tr>
<tr>
<td>Gain factor $g_i$</td>
<td>20</td>
<td>5 bits on subframe</td>
</tr>
<tr>
<td>Vector of stochastic code book $C_s$</td>
<td>36</td>
<td>9 bits on subframe</td>
</tr>
<tr>
<td>Gain factor $g_s$</td>
<td>20</td>
<td>5 bits on subframe</td>
</tr>
<tr>
<td>Synchronization</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Forward error correction</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Reserve</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>In total on a segment (30 ms)</td>
<td>144</td>
<td></td>
</tr>
</tbody>
</table>

Codes with Multi Band Excitation (MBE)

Among the most widespread algorithms basing on MBE models should be named:
- IMBE (Improve Multi Band Excitation) 6.4 kbit/s, developed DVSi (Digital Voice System Inc.) and accepted as the standard for Inmarsat-M in 1990;
- AMBE (Advanced Multi Band Excitation) 4.8 kbit/s, DVSi and accepted as the standard for Inmarsat mini-M in 1994;

The MBE-model has appeared rather recently (it is developed Massachusetts Institute of Technology, USA) and has much smaller distribution, than its "competitors" - models of a linear prediction with every possible excitation.

MBE-model of speech production

In classical model of speech production for voice coders with a speed 4+6 kbit/s and is lower, samples of a speech signal shared on two classes (according to the nature of a excitation signal). For voiced speech the excitation signal represented periodic sequence of pulses, and for unvoiced - realization of white noise. Thus, the accepted decision "voiced/unvoiced" concerned to all speech segments and defined a method of formation of all segment. Thus, the accepted decision "voiced/unvoiced" concerned to a speech segment as a whole and defined a method of formation for all segment. The basic difference of MBE-speech model consists in the approach to formation of excitation. The spectrum of an excitation signal shares on not overlapped frequency bands (see figure 4), and the decision "voiced/unvoiced" is accepted in each band separately.

Thus, the signal of excitation will consist simultaneously and from voiced and from unvoiced frequency component. All this allows to increase a degree of freedom in excitation modeling, and to receive higher quality of speech synthesis. Besides it is reflected in the best stability to background noise.

In figure 4 the situation in which the MBE-model reflects a signal with the mean spectrum better is shown.

So, the classical model classifies a speech segment as voiced and uses periodic exciting sequence. The MBE-model (let it has 10 frequency strips) classifies bands 4, 5, 9 and 10 as unvoiced, and uses for them noise excitation (and for the others pulse excitation).

The band is classified as voiced, if the attitude of periodic signal components energy to noise components energy is high, and unvoiced if this attitude is low.

The codec using MBE-model, estimates amplitudes of a signal in each band, transfers these data to the reception side where with help fast Fourier transformation forms the mixed spectrum.

Conclusion

Feature of voice transmission in digital communication systems using a method of message source coding is transfer on a communication channel not the speech signal, but parameters of it model. Development of methods of excitation signal representation has an influence on development of the theory and practice of construction of digital speech transmission system.

In given article in view of historical dynamics are considered the most popular methods of formation of an excitation signal, used in digital communication systems. The share of information expenses for coding of an excitation signal in the appropriate systems is given.

The order of a statement of a material and the given numerical parameters allow looking after the basic tendencies of development of algorithms of speech signal modeling. The account of features and tendencies of an excitation signal modeling allows defining perspective directions of perfection of speech analysis and synthesis devices based on a linear prediction method.
Figure 4. Illustration of MBE-model of speech production