УДК 621.39

O. M. Tkachenko, Cand. Sc. (Eng.), Assoc Prof.;
L. V. Krupelnytskiy, Cand. Sc. (Eng.), Assoc. Prof.; S. V. Hruschak

ADAPTIVE VOICE ACTIVITY DETECTION IN DELTA COMPRESSION OF SPEECH SIGNALS

The paper presents the developed method of Delta-compression with the prediction of the subsequent index of the quantized value in the code book and the improved method of the voice activity detection described in the recommendation G.729, Annex. Experimental investigation of the proposed method was performed.

Key words: *compression of speech signals, linear prediction coefficients, linear spectral frequencies, vector quantization, voice activity detection.*

Current importance of the research

The main problem of speech transmission over digital channels is the amount of information to be transferred per time unit in order to provide qualitative voice communication. Speech compression reduces the transmitted data volumes and hardware costs.

This paper considers the process of speech signal compression using LPC vocoder and proposes a number of approaches to further reduction of data volumes.

Introduction

Overwhelming majority of speech compression algorithms use a linear prediction model for signal description and encoding. In accordance with this model spectral information about a signal is described by 10 linear prediction coefficients (LPC). For more efficient quantization and further interpolation LPC are, as a rule, converted into linear spectral pairs (LSP) and quantized using code books.

LSP obtained as a result of speech encoding using the linear prediction method have certain dependencies of the elements both inside the frame and between adjacent frames, which is the evidence of the high level of excessiveness when these coefficients are used [1].

Taking into account correlation between the coefficients it is possible to reduce the volume of data required to describe the speech signal parameters and to reduce, in this ways, the requirements to the capacity of the information transmission channel. This research considers a parametric system of speech compression that takes into account correlation between the coefficients inside the frame and between the successive frames making it possible to achieve a higher degree of compression.

The next stage in the reduction of data volume to be transmitted via communication channel is application of discontinuous transmission (DTX) – disconnection of the transmitter or transmission of the comfort noise (CN) to the receiving side during pauses in the conversation. Such approach is used, for instance, in GSM network. In this case the main task is voice activity detection (VAD). Most approaches to the voice activity detection are based on the energy thresholds, detection of the main tone period, spectral analysis, zero crossing frequencies or on the combination of several methods. The complexity of this task consists in the following: it is impossible to achieve constant high accuracy of the signal frame vocalization detection as most of the algorithms are based on a certain threshold value that is a fixed one or is calculated from the signal on noise segments. E. g., when for VAD the method of least squares is employed, noise segments are used in the filter with finite pulse characteristic for its training. But in order to achieve high accuracy of the VAD algorithm in the cases when background noise is not stationary or speech signal is mainly a vocalized one, the threshold value must be constantly corrected irrespective of the frame type. For this a method is

required that could be effectively adapted to the background noise changes.

This paper improves the method for voice activity detection proposed by Internarional Association ITU-T – G.729 Annex B [3] developed for multimedia and IP telephony. This method provides adaptation to the high level of background noise and uses the data obtained in the coding process, which reduces its computational complexity. But, in spite of this, it has low speed compared with other standards (ETSI AMR, statistical models), especially for low signal / noise ratio. That is why certain changes to this method are proposed in order to improve the speed and its investigation is conducted using the proposed procedures.

Taking into account correlation between the coefficients

Correlation between the coefficients of one frame is taken into account through the use of vector quantization of LSP coefficients. In vector quantization a set of LSP coefficients is considered as one vector for which the closest quantized value is searched in a special table – a vector code book [4].

For taking into account correlation between successive frames work [4] proposes a method for ordering the vectors in accordance with the majorization relation, which allows making a transition to transmission of the difference between the indices of successive frames. A shortcoming of this method is a considerable delay (100 ms) required to obtain a qualitative signal when the amount of information for description of the speech signal parameters is reduced from 24 to 20 bit, which is impermissible in some communication systems. Taking this into account, Delta compression method with the subsequent count prediction is proposed, which makes it possible to get shorter delay (20 ms) with inconsiderable reduction of the index transmission accuracy.

To reduce Delta window between the indices of the adjacent frames with a shorter delay as compared with the method proposed in [5], frame index prediction on the basis of preceding values is proposed. Prediction is performed by creating an extrapolating function for the indices of several frames at the receiving and transmitting sides. When compression is performed delta between the predicted and real values is also transmitted to the receiving side. The proposed method is illustrated in fig.1.

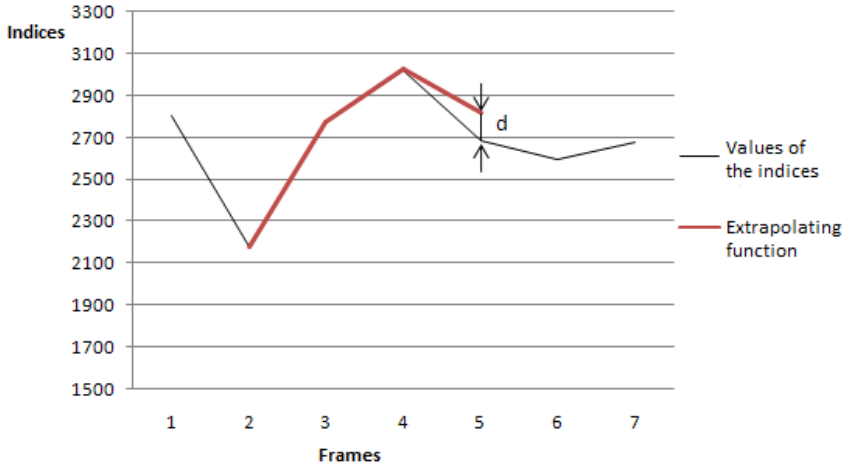

Fig. 1. An example of how the method of compression with extrapolation works

Extrapolation function is built using the method of least squares. This method consists in the following: function g for the experimental data description is built as linear combination M of the basis functions F_j [6]:

$$g = \sum_{i=0}^{M-1} c_j F_j(x). \qquad (1)$$

Here the coefficients c_j are chosen so that criterion C_{MLS} – the sum of the squared deviations of the extrapolating function from the experimental values – would be minimal.

$$C_{нмк} = \sum_{i=1}^{n} (y_i - g_i)^2 \to \min C, \tag{2}$$

where y_i is the experimental function.

Alternative methods were also investigated – the method of the least modules (MLM) and Chebyshev method. For MLM the sum of the deviation of modules is a minimization criterion (3):

$$C_{мнм} = \sum_{i=1}^{n} |y_i - g_i|. \tag{3}$$

For Chebyshev method maximal deviation is a minimization criterion [6]:

$$C_{мч} = \max |y_i - g_i|. \tag{4}$$

However, only the method of least squares makes it possible to find he best c_j within a finite number of operations reducing the problem to the solution of the linear equation system. This means that it is the simplest method for calculations and it gives more accurate description of the experimental function.

Introduction of the delay allows correcting the values of the indices to be transmitted. This is necessary in the situations when the prediction proves to be inaccurate (fig. 2), e.g. in the given case we must predict the index of the fifth frame on the basis of the preceding ones – the second, the third and the fourth. However, the prediction has made the result only worse: the value of delta increased.

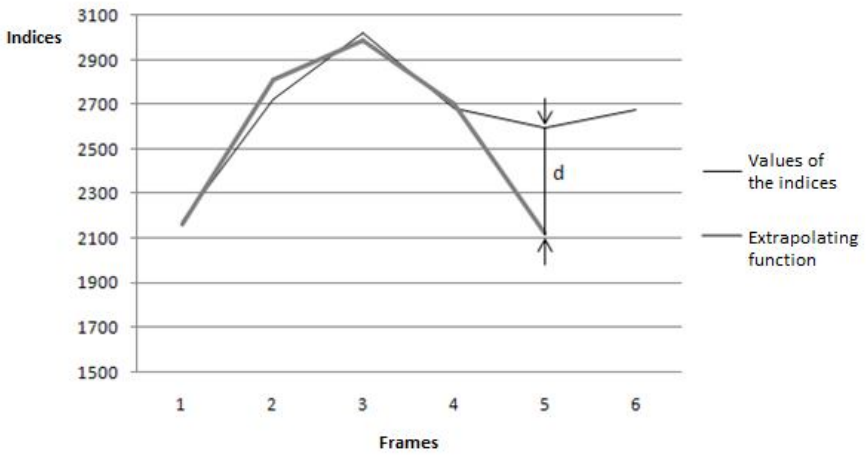

Fig.2. An example of inaccurate prediction

In the cases when during transmission of the *i*-th frame inequality (5) holds

$$\Delta_{i+1} > S_{w,} \tag{5}$$

where Δ_{i+1} is the value of delta for *i*+1 frame, S_w – the size of the window, the following algorithm is used:

1. Approximation of the experimental function values is performed taking into account the following values of the indices *i+1, i+2, ..., i+k*, where k is the value of the delay in the frames. For approximation the method of least squares is also used.

2. Correction of the *i*-th index is performed according to the approximating function in order to receive the best prediction for the subsequent frame.

3. Prediction of the *i+1* frame is performed on the basis of the preceding frames *i, i-1, i-2, ..., i-p*, where *p* is the number of frames to create a prediction that is determined by the order of the prediction model.

4. If the difference between the real signal and the predicted one goes beyond the limits of the window, the correction is repeated: it is necessary to go back to point 2. In the extreme case when for a maximally permissible deviation the prediction proves to be inaccurate, it is necessary to pass to

point 5.

5. The difference between the corrected value and the the predicted one, taking into account the values of indices for *i-1, i-2, ..., i-(p+1)* frames, is transmitted to the communication channel.

The result of the algorithm implementation is shown in fig. 3. After correcting the index value for the fourth frame the prediction is advancing successfully – the value of delta remains inconsiderable.

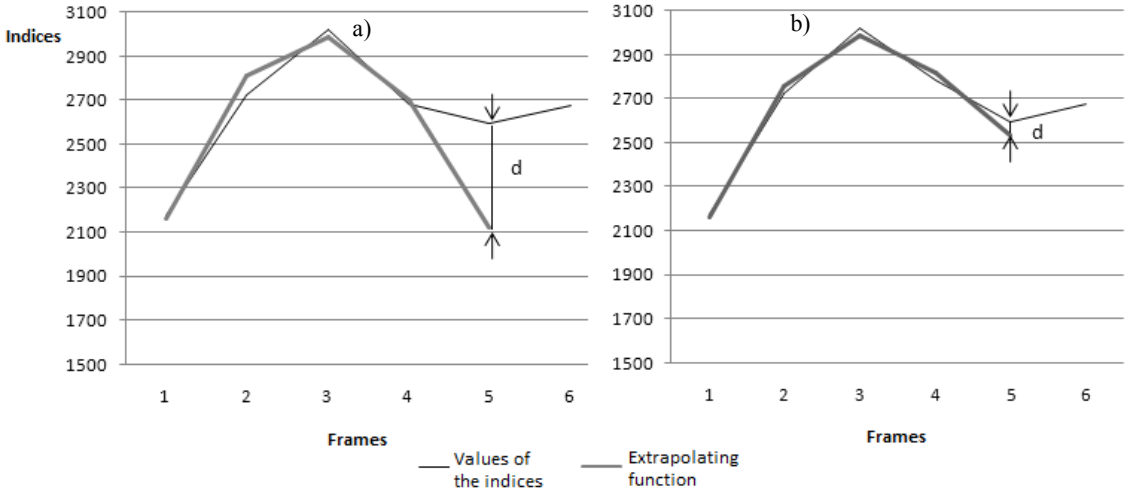

Fig. 3. Plots of the experimental and extrapolated values of the indices:
a) before correction;
b) after correction.

Improvement of the voice activity detection algorithm

Recommendation ITU-T G.729 Annex B is developed as a supplement to the vocoder G.729. Decisions about vocalization are taken for the frames with the duration of 10 ms on the basis of the following characteristics:

1. Differential energy of the signal for the entire spectrum range – ΔE_f;
2. Differential energy of the signal for the low-frequency spectrum (0 – 1 kHz) – ΔE_l;
3. Differential frequency of zero crossings – ΔZCR;
4. Spectral distortion – ΔLSF.

The diagram of this VAD procedure is given in fig. 4.

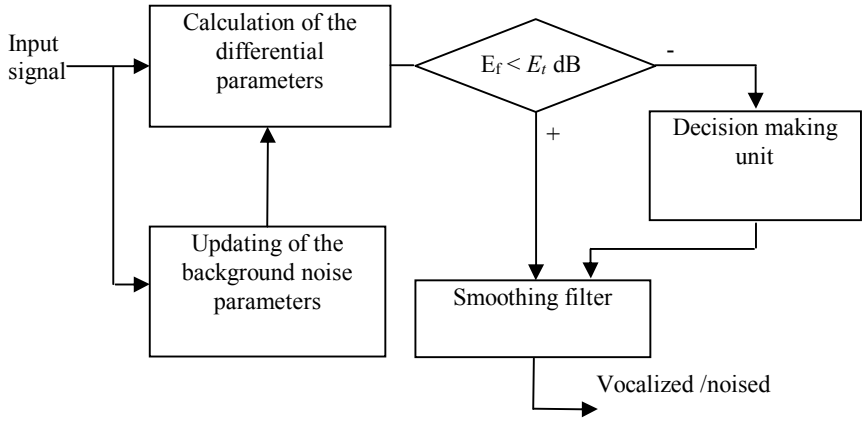

Fig. 4. Block-diagram of the ITU-T G.729 Annex B algorithm

Input parameters are obtained from the speech signal using a first-order autoregressive model if the condition ΔEf < Et (where Et is a pre-set threshold value) is fulfilled. However, these parameters involve high computational complexity as well as the use of zero crossing frequencies and energy threshold, which causes errors under high noisiness of the signal.

To improve the quality of the proposed method, acceleration of the spectral Stroke evaluation is

proposed by means of smoothing using MLS. This simplifies the spectrum flatness calculation that is a key factor for a signal frame vocalization detection (for a non-vocalized fragment the spectral Stroke will be flatter). An example of such smoothing of the spectral stroke is presented in fig. 5.

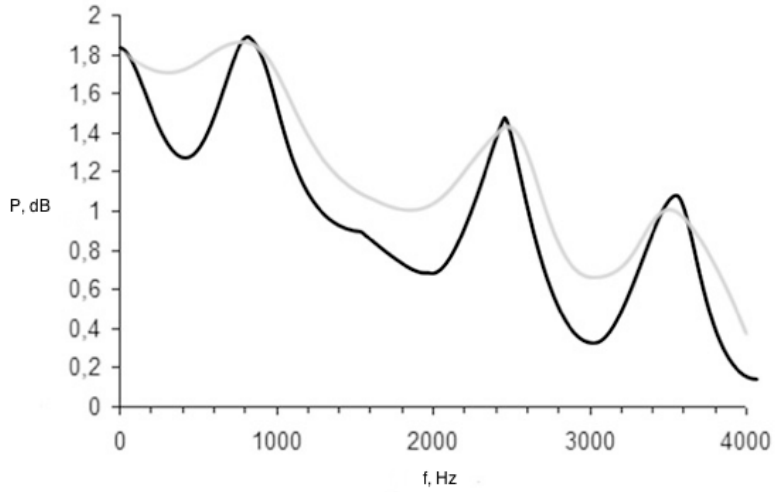

Fig. 5. Example of smoothing the spectral stroke of a signal

Hence, this makes it possible to reduce the number of computations, because the spectrum flatness is calculated according to several points, as well as to increase the algorithm accuracy, especially in the low-frequency band of the signal.

In order to calculate maximal threshold value of the signal vocalization, a new approach is used – geometrically adaptive energy threshold proposed by Öser and Tanyer [7]. This approach is based on the probabilistic distribution of the amplitudes and enables adaptive setting of the threshold value using not only noise signals but also vocalized ones. The method is stable to non-stationary noise but sometimes could give an incorrect result under short noise spikes. Due to simplicity of the adaptive energy threshold calculation, complexity of the VAD algorithm is also reduced.

Delta compression method of investigation

Testing of the method for the enhanced search of the vectors was performed on the basis of the phonetic material consisting from two phonetically complete texts [4]. When the delta compression method with the prediction of the index value in the code book is used, prediction for the subsequent frame is performed by creating the extrapolating function for the indices of several frames. Several methods of the extrapolating function construction were investigated: the method of least squares (MLS), Chebyshev method and the method of least modules (MLM).

Dispersion of delta D probability distribution between successive frames and spectral distortion SD were used as evaluation criteria of the methods. Evaluation results are presented in table 1.

Table 1

Results of investigation of the extrapolation methods for a code book with the size of 4096 vectors

Name of the method	D	Several SD, dB	Percentage of frames for which SD>2 dB, %	Percentage of frames for which SD > 4 dB, %
MLS	21531	0,882	1,78	0
Chebyshev method	31095	0,912	1,96	0
MLM	141950	0,956	2,15	0

As it is evident from table 1, MLS is the best method according to the indices of spectral distribution and spectral distortion. For MLM the number of frames with spectral distortion in the range of 2 – 4 dB exceeds the permissible value of 2 dB adopted for the systems of low-speed speech compression.

Investigation of the method for voice activity detection

A commonly adopted practice of VAD method quality evaluation is the use of one from the two types of tests – objective or subjective ones. Subjective methods are those where the algorithm work is evaluated with the help of listeners' estimates. However, such scheme does not make it possible to determine the number of inaccurate operations for high-level noise as this will be expressed in the increased amount of data in the channel and not in the resulting speech quality. Therefore, the objective method of VAD evaluation has been chosen. TIMIT speech base was used for the tests, with random noise being added to its recordings. Then the records were divided into frames which were marked manually as vocalized and noise frames, i. e. in fact, comparison with the ideal variant was conducted. These files were processed using different algorithms of voice activity detection – the proposed, the original, ETSI AMR 1 and ETSI AMR. After that the number of fragments with speech clipping and where noise was recognized as speech was calculated. Testing results are given in fig. 2.

Table 2

Quality testing results

Name of the method	Frames with speech, %	Noise frames recognized as vocalized ones, %
The improved ITU-T G.729B method	83%	2%
ITU-T G.729B	68%	5%
AMR 2	87%	0,3%
AMR 1	76%	5,5%

Testing results have shown that the improved ITU-T G.729B method makes it possible to achieve higher quality of speech recognition on noise frames than the ordinary one, but lower quality than the method where high-order statistics was used which, however, requires more computational efforts. If we take into account the fact that in two-side conversation each of the speakers talks for about 40 % of time [2], then this method could reduce the communication channel load to 60 %, though in real conditions it does not exceed 45 %.

To investigate the speed of the proposed improved method of voice activity detection, materials from TIMIT set were used as this enables comparison of the obtained results with the foreign analogs. The results of comparing the improved ITU-T G.729 Annex B method with the existing methods for analyzing different-size units are presented in table 3.

As it is evident from the table, the improved method works faster than the original one and the method based on high-order statistics. Method with the application of adaptive high-energy levels is the fastest but also the least accurate one.

Table 3

Average quantity of operations for the proposed and the existing methods for different-size units

Name of the method	The number of mathematical operations for the given size of the analyzed unit, ms			
	10	20	50	100
The improved method ITU-T G.729B	$1{,}7\cdot10^4$	$2{,}4\cdot10^4$	$5{,}2\cdot10^4$	$14{,}3\cdot10^4$
ITU-T G.729B	$2{,}2\cdot10^4$	$2{,}7\cdot10^4$	$6{,}1\cdot10^4$	$18{,}6\cdot10^4$
AMR 2	$2{,}9\cdot10^4$	$2{,}6\cdot10^4$	$4{,}8\cdot10^4$	$9{,}8\cdot10^4$
AMR 1	$1{,}6\cdot10^4$	$2{,}0\cdot10^4$	$3{,}9\cdot10^4$	$7{,}2\cdot10^4$

Conclusions

Investigation of the method of delta compression with prediction has shown that its application makes it possible to reduce the number of bits for transmission of speech parameters from 24 to 20 with one-frame delay and inconsiderable (SD – 0,882 dB) signal distortion. Application of the discrete transfer of packets enables further channel load reduction to 45%. The proposed improved ITU-T G.729B method makes it possible to achieve higher quality of noise frame recognition than the ordinary method with lower complexity of computations than ETSI AMR 2.

REFERENCES

1. Chu W. C. Speech Coding Algoritms – Foundation and Evolution of Standardized Coders. / W. C. Chu. New Jersey:Wiley. – 2003. – 553 p.

2. Ramirez J. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding. / J. Ramirez, J. M. Gorriz and J. C. Segura. Vienna:I-Tech Education and Publishing. – 2007. – 460 p.

3. Benyassine A. ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application. / A. Benyassine, E. Shlomot, and H.-Y. Su // IEEE Communications. – 1997. – № 35. – P. 64 – 73.

4. Ткаченко О. М. LSF-вокодер на основі векторного квантування / О. М. Ткаченко, Н. О. Біліченко, О. Д. Феферман, С. В. Хрущак // Реєстрація, зберігання і обробка даних. – 2007. – № 1. – С. 35 – 41.

5. Ткаченко О. М. Метод дельта-ущільнення мовних сигналів / О. М. Ткаченко, О. Д. Феферман, С. В. Хрущак // Інформаційні технології та комп'ютерна інженерія. – 2008. – № 1(11). – С. 8 – 13.

6. Коваленко И.Н. Теория вероятностей и математическая статистика: Учебное пособие. / И. Н. Коваленко, А. А. Филиппова. – М.: Высшая школа, 1982. – 256 с.

7. Özer H. A geometric algorithm for voice activity detection in nonstationary Gaussian noise / H. Özer and S. G. Tanyer // EUSIPCO, – 1998. – № 1. – P. 23 – 26.

Tkachenko Oleksandr **–** Assoc. Prof. of the Computer Engineering Department.

Krupelnytskiy Leonid **–** Assoc. Prof. of the Computer Engineering Department.

Hruscak Sergiy — Post-graduate student of the Computer Engineering Department.
Vinnytsia National Technical University.