Development and Research of VAD-Based Speech Signal Segmentation Algorithms

Oleksandr Tymchenko[1][0000-0001-6315-9375], Bohdana Havrysh[2][0000-0003-3213-9747], Aneta Poniszewska-Maranda[3][0000-0001-7596-0813] ,Bohdan Kovalskyi[2][0000-0002-5519-0759] Oleksandr O. Tymchenko[2][0000-0003-2774-2138] and Kateryna Havrysh[4][0000-0003-4155-8759]

[1] University of Warmia and Mazury Olsztyn, Poland
[2] Ukrainian Academy of Printing, Lviv, Ukraine
[3]Technical University of Lodz, Lodz, Poland
[4] IT Step University, Lviv, Ukraine
o_tymch@ukr.net
dana.havrysh@gmail.com, bkovalskyy@ukr.net
olexandr.tymch@gmail.com
aneta.poniszewska-maranda@p.lodz.pl
gavrysh.kateryna@gmail.com

Abstract. The method of a speech signal segmentation developed during the work by application of the VAD detector uses a spectrum of power of fragments (packets) of a speech signal unlike the other known examples. A discrete Fourier transform with a small number of samples (maximum 160) is used to calculate the spectrum. The developed method allows not only to solve the traditional problem of VAD – the data rate reduction, but also to perform the speech signals separation and segmentation into individual fragments. Examples of such segmentation and determination of vocalized and non-vocalized areas of speech signals boundaries in the data network are given, which can be used to build phonemic vocoders in automated speech processing and recognition systems.

Keywords: segmentation, speech signal, communication channel, speech data.

1 Introduction

Packet data networks have occupied and hold the leading position among telecommunication networks, in which they are facilitated by the computer networks development and the Internet. One of the main types of packet traffic is a multimedia traffic, which has a significant place in the language signal. Various encoders [1] are used to provide high-quality voice traffic [1, 3], which simultaneously compress signals to reduce network congestion. An effective means of further compression ratio enhancing is the use of the most current language codecs of Voice Activity Detector (VAD) technologies [2, 4]. Even greater increase in the compression degree is achieved by the methods of speech fragments separation and segmentation, ie as a result of the transition to phonemic and semi-phonemic vocoders.Typically, none of the low-speed

voice encoder implementation can do without the use of VAD (Voice Activity Detector) technology. The process of identifying or absence of voice activity is not a new task, different methods of its implementation have been and are still being used (eg GSM encoders, different methods of speech recognition, etc.). A well-known problem with VAD synthesis using voice signal encoders for VoIP networks is to correctly identify language pauses against a background of intense acoustic noise (office, street, car, etc.). However, the use of VAD can significantly save bandwidth [4] and therefore congestion of network channels.

2 The research of existing VAD technologies capabilities

VAD provides the ability to pre-process the speech signal before it is fed to the encoder. In the first approximation, the following types of linguistic fragments can be distinguished: vocalized, unvocalized, transitional, and pauses. When a language is processed into the digital form, ie in the form of a sequence of numbers, each signal type having the same duration and quality requires a different number of bits for encoding and transmission. Therefore, the transmission rate of different fragments of speech signals may also be different. Thus, an important conclusion can be made here: the linguistic data transmission in each direction of the duplex channel should be considered as the transmission of asynchronous logically independent fragments of digital sequences. These sequences (transactions) contain batch (datagram) synchronization inside a transaction filled with packets of different lengths [5].

The VAD detector must be sensitive and responsive in order to avoid the loss of the word beginnings when switching from the pause to the active speech fragment. At the same time the VAD detector should not be triggered by background noise [4, 7].

Generally, the VAD goal is to estimate the value of a particular input parameter (eg, level, power, etc.) and, if it exceeds a certain threshold, then such a packet will be transmitted. This slightly increases the delay in the speech signal processing in the encoder, but it can be minimized by creating coders that work with packets (datagrams) of the readings.

In the encoder analysis with the fast use of the C_{code} language. (Bit / s), the signal is divided into individual fragments (usually quasi-stationary sections), duration T_{fragm} from 2 to 50 ms and is in the input block used with N difference, uses the usual information frame about $V_{m.k} = T_{fragm} \times C_{code}$ (bits).

No matter what are the details of the implementation, the main criterion for evaluating the encoder is the high quality of speech reproduction at a low output speed of digital C_{code} output. Especially of an output with minimum requirements for the digital signal processor resources and minimal delay [6].

technology can be combined with a wide variety of language encoders.

1. There is a method of detecting voice activity based on finding the formant. Although formants carry the basic spectral information about the speech signal, in the case of unvocalized areas their localization is unreliable and segmentation is ambiguous because it is lost in the noise [3].

2. In a number of works the spectral characteristic of noise is estimated and on the basis of it the speech signal from a mix of a signal and noise is separated. The GSM standard adopted a VAD circuit with frequency domain processing [8, 16]. A block diagram of such a VAD system operation is shown in Fig. 1. The essence of its work is based on differences of speech and noise spectral characteristics. Background noise is considered to be stationary over a relatively long period of time, and its spectrum is slow to change over time. Therefore, VAD estimates the spectral deviations of the input sequence from the background noise spectrum. This operation is performed by an inverse filter whose coefficients change according to the input action. In the presence of a speech signal and noise input, an inverse filter suppresses the noise components and, in general, reduces its intensity. The energy of the signal + noise sum at the output of the inverse filter is compared with the threshold, which is variable and is estimated during the periods of action at the input of the noise itself. This threshold is higher than the noise signal energy level. Exceeding this level is a determining criterion for the presence of voice activity input. Because these parameters (coefficients and thresholds) are used by the VAD detector to detect the language, it is not for the VAD to decide at this stage of the analysis, as the threshold may vary [18].

This decision is made by a secondary VAD based on a comparison of the envelope spectra in successive periods of time. If they are similar or close for a relatively long time, then it is assumed that noise is applied at the detector input, then the filter coefficients and the noise threshold can be varied, ie adapted to the current level and spectral characteristics of the input noise [9, 17].

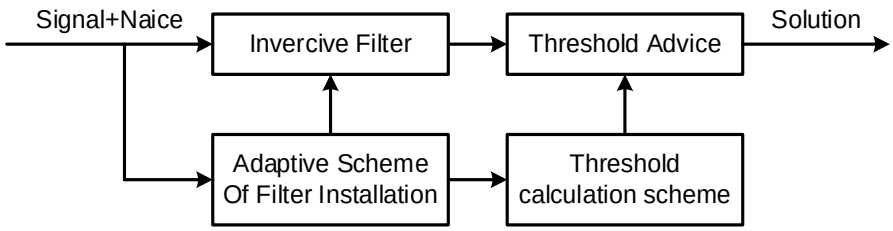


Fig. 1. VAD structural diagram on spectral characteristics of noise

A clear disadvantage of this scheme VAD is the "relatively long period of time" for which voice activity is decided [10, 12]. In addition, if the noise is non-stationary it is almost impossible to segment the speech signal with such a scheme.

3 The algorithm of speech signal segmentation is offered

The main idea behind the proposed VAD-based speech signal segmentation algorithm is the linear processing of the speech fragments and the rejection of fragments where there is no voice activity (ie useful information).

The input parameters for the algorithm are the minimum length of language data - Mframe considered useful (number of packets and their duration), maximum pause time in the composition or word Eframelength, ie "error" VAD (obviously, this error can take zero if the VAD system responds to the lowest possible signal values).

The program code for the language segmentation algorithm using VAD is as follows.

```
Mframelength = 5...X;
Eframelength = 0...Y;
L = 0; - counter
ArrayList s;
ArrayList f;
int begin = 0;
int a = 0;
while (L < Plength)
  if (p[L] = true)   – voice activity indication
begin = L;
    a++;
  else
    if (a > Mframelength )
      s.Add( begin );
      f.Add( L – 1 );
L++;
L = 0;
if (Eframelength > 0)
while ( L < s length – 1 )
  if ( s[L+1] < e[L] )
    s.RemoveAt([L+1]);
    f.RemoveAt([L]);
L++;
return s,f;
```

After the selection, the language fragments are detailed and processed according to conventional coding algorithms (according to ITU-T Recommendation).

4 Methods of experimental research

The proposed VAD is based on a discrete Fourier transform (DFT):

$$a_k = \frac{2}{N}\sum_{i=1}^{N} y_i \cos\left(\frac{2\pi ki}{N}\right), b_k = \frac{2}{N}\sum_{i=1}^{N} y_i \sin\left(\frac{2\pi ki}{N}\right), S_k = \sqrt{a_k^{\,2} + b_k^{\,2}} \quad (1)$$

Depending on the selected packet length, we choose the number of spectral components (from 1.2 to $N/2$, where N is the packet length).

The band in the frequency spectrum (ΔS that is [0-1] by default, determined relative to the number of harmonics) is selected as the main parameter of the VAD block.

To study the effectiveness of the proposed algorithm, a simulation of its operation using real language signals was conducted. The scheme of the study is shown in Fig. 2.

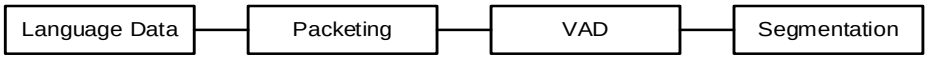


Fig. 2. Scheme of language segmentation algorithms research

The scheme includes:

- "Voice Data" supports the download of an arbitrary speech signal file in Pulse Code Modulation (PCM) format. At the output of this block we get an array of speech signal samples with representation in integer or floating point format (selected from the criteria of data representation accuracy), the number of which is determined by the sampling frequency of the input signal and the packet length: S_k , $k = 0..n$ – output signal.
- The Packetization block splits the language data into fragments. Usually the length of these fragments corresponds to the stationary period of the speech signal. Using the data [13, 14] we assume that this value is no more than 20 ms (this parameter P is specified in the counts and can take values from 1 to the total number of input counts n).

The packeting algorithm is described as follows:

$Sp_{i,j} = S_k$, $i = 0..P$, $j = 0..k/P$ – the counting value (i) in a given packet (j).

After a cyclic change of i and j within the input data duration, we receive a generated packets stream of a given duration (at a sampling rate of 8000 Hz and a packet duration of 20 ms we have 160 samples in the packet).

- the Voice Activity Detection Unit may use an arbitrary algorithm. The VAD level and DFT based VAD were used in the simulation process [15].

The VAD system replaces the packets with "zero" by the signal level (ie, all samples at the block output are equal to 0), if 80% of the packet samples are less than the specified threshold. The selected threshold (delta) is a given value of quantization levels (steps). This value can be changed from 0 to 127 quantization levels with a maximum value of signal amplitude 255 (for eight-bit quantization)

Signal level VAD algorithm:

```
L = 0 – %  less than the delta threshold counter count-
down
for (i,j)
i = 0..P, j = 0..k/P;
if(abs(SPi,j)<delta)
L++;.
if ( L > 0.8P )
Sj = 0;
```

That is, if the selected criterion of "informativeness" is not fulfilled, then the package must be zeroed.

When using a VAD system based on DFT, the DFT samples of the speech fragment are calculated:

$SP_i[]$– spectrum amplitude package, where j=0..n;

$S_i[]$– packet-matching in the language input stream.

One step of the algorithm is as follows:

```
for(j) – for all packages j=0..n:
if (max (SPj ) ∈ ΔS)) Sj[]=0;
```

that is, if the selected criterion of “informativeness” is not fulfilled (the maximum of the amplitude spectrum is in a given band), then the packet is nullified, ie it is concluded that the packet does not carry any useful language load. It is advisable to choose the band ΔS in the range from 0 to 100 Hz, since the frequency of the pitch of the speech signal is always above 200 Hz [1, 20-21].

The "Segmentation" block performs the operation of combining packets with prescreening of areas where there is no linguistic activity, according to the above algorithm for segmentation of language using VAD.

5 Results of experimental studies

Two signals of up to 2 seconds in length were selected to study the segmentation method, which correspond to the language fragments (file 1 and file 2). Segmentation was performed using VAD by level with a threshold value of 0.0625 (relative to 1) and 0.125 within the limits shown in Fig. 3 and 4 (file 1). The segmentation of the speech signal using VAD based on DFT (file 1) is shown in Fig. 5. The speech signal segmentation using VAD based on DFT (file 1) is shown in Fig. 5.

File 1 from the 1.25-second language stream (10,000 samples) was selected for processing. The 5760 samples fragment (from 2240th to 8000th count) was highlighted as a result of the VAD level application with a threshold value of 0.0625. The bypass of the input language signal and highlighted fragment are represented in Fig. 3. The research result is highlighted with the blue lines, which practically corresponds to the relevant linguistic information.

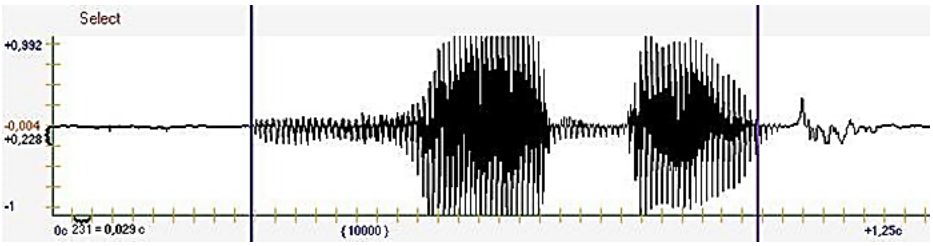


Fig. 3. Language data segmentation (file 1), VAD level 0,0625.

To study the language segmentation operation algorithm with VAD by signal level, a gradual increase in the threshold value was performed. As a result, upon reaching the threshold value of 0.125, two segments of 1600 and 1220 counts were obtained, respectively. The selected fragments correspond to the loud sounds "a", and at the beginning of the second fragment there is a deaf sound "b. The results are presented in Fig. 4, where the vertical lines indicate two sections: the first (a) is in the range from the 4000th to the 5600th reference, the second (b) displays from the 6560th to the 7840th reference.

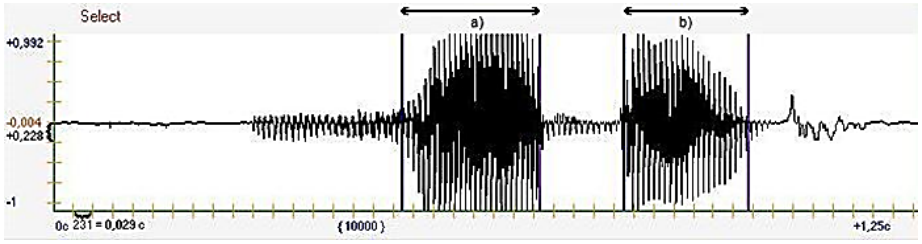


Fig. 4. Language data segmentation (file 1), VAD level 0.125

Applying DFT-based VAD segmentation method to file 1, three segments were obtained (Fig. 5). For greater clarity, they are presented separately: the first fragment in the range from 2560 to 3360 (Fig.6a), the next fragment in the range 3680 – 6080 (Fig.6b), and the last frame from 6400 to 8160 reference (Fig.6c). Thus, the use of VAD on the basis of DFT allowed to distinguish a fragment with linguistic activity, which at VAD level was missed.

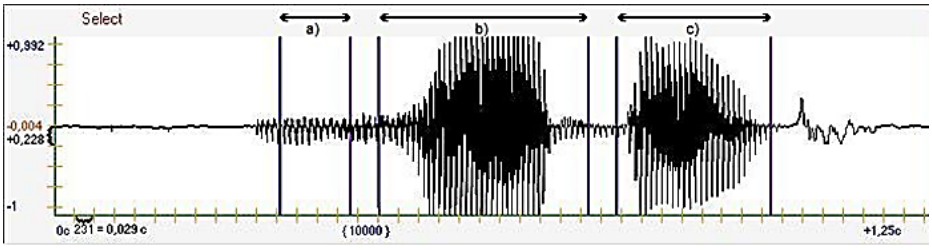
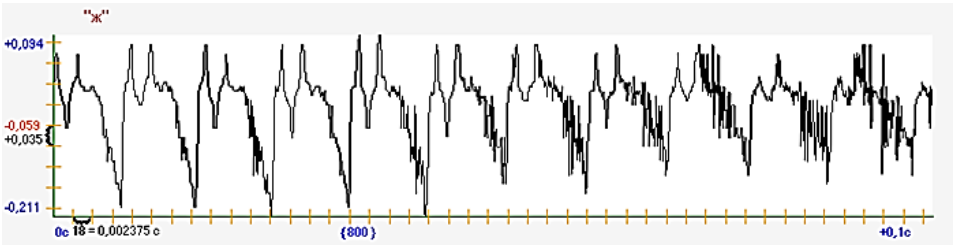
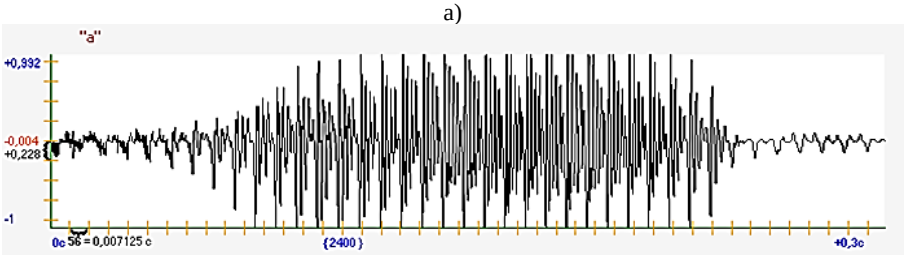


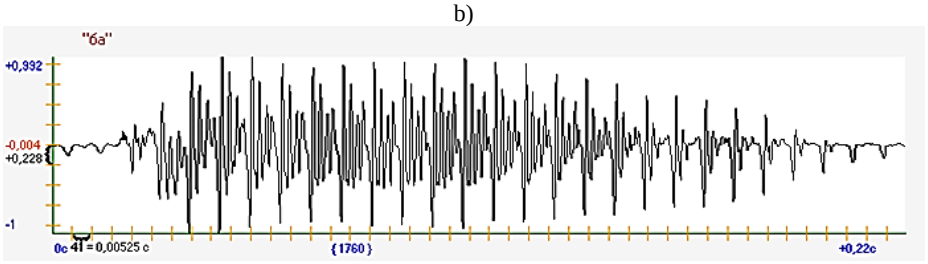
Fig. 5. Language data segmentation - VAD (file 1), based on DFP



a)



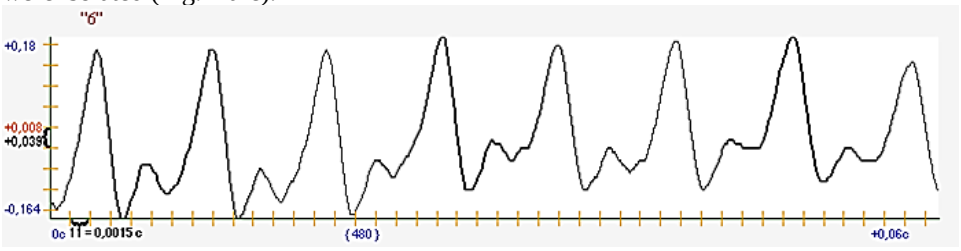
b)



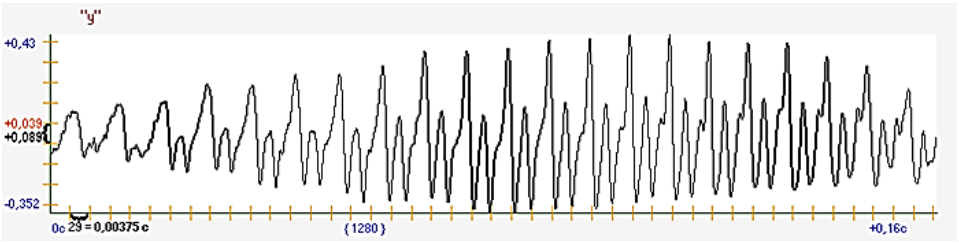
c)

Fig. 6. Segmented data - file 1 (the scale on the abscissa axis in the figures is different). VAD based on DFT.

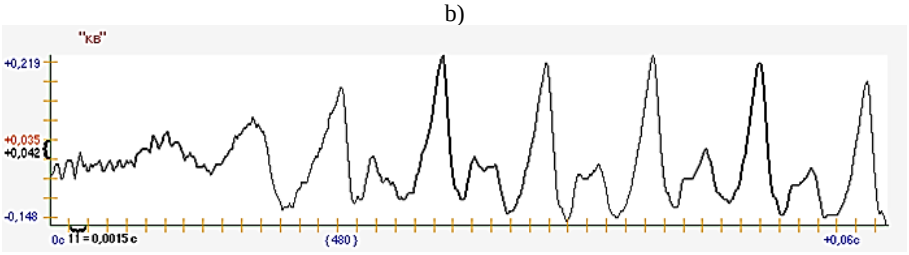
In the same way, the segmentation of the linguistic data presented in file 2 was carried out. As a result of the use of VAD on the basis of DFT, 5 linguistic fragments were isolated (Fig. 7 a-e).



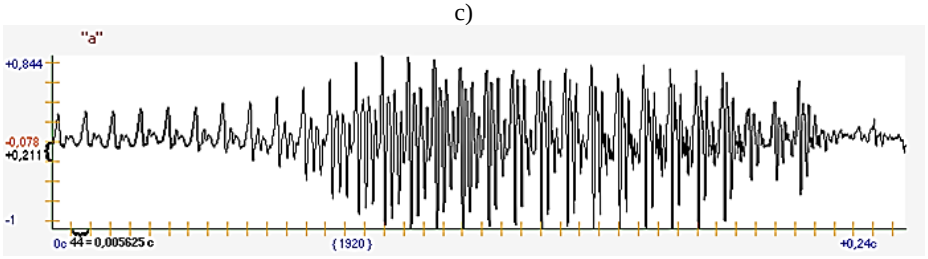
a)



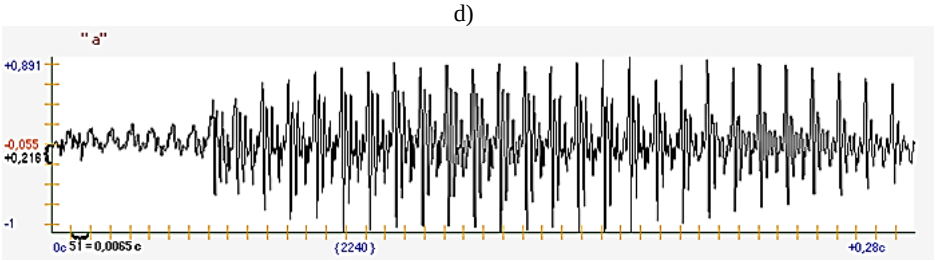
b)



c)



d)



e)

Fig. 7. Segmented data - file 2 (the scale on the abscissa axis in Fig.6.a-e is different). Used VAD based on DFT

6 Conclusion

The analysis of the research results showed that the developed segmentation algorithm using VAD with DFT gives almost error-free division of the speech flow into words, and even into syllables and letters depending on the speaker's intonation. Moreover, depending on the intonation of the speaker - even into syllables and letters. Also, raising the VAD threshold to the level provides a virtually error-free selection of vocalized language fragments. The main drawback of the VAD algorithm based on

DFT is the lack of sensitivity when using signals in the [300..3400] Hz range, as a result of which segmentation into letters is rarely achieved, unlike signals in the [0..3400] Hz range. However, the proposed VAD technique can be effectively used in language recognition, since the first DFT harmonics provide additional information about formats, which can be used to detail individual letters or syllables.

Comparative analysis of test signals using objective quality assessment (PESQ) shows that the intelligibility of the speech signal remains practically at the same level (3.7-4.5). A score of 3.7 corresponds to the fragments of the language where the low-power packets were zeroed.

With respect to the gain in compression and subsequent transmission of the variable speed encoded signals using VAD, a gain of 1.5-2 times (34 / 75 frames and 73 / 150) can be obtained if the transmission of empty packets is stopped or transmitted a special short code sequence.

Acknowledgments. The authors are appreciative to colleagues for their support and appropriate suggestions, which allowed to improve the materials of the article.

References

1. J. Cai, "Noise estimation using an MVDR-like approach for acoustic signal enhancement," IET International Conference on Information and Communications Technologies (IETICT 2013), Beijing, China, 2013, pp. 192-200, doi: 10.1049/cp.2013.0053.
2. S. Ou, W. Liu, S. Shen and Y. Gao, "Two methods for estimating noise amplitude spectral in non-stationary environments," 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, 2016, pp. 969-973, doi: 10.1109/CISP-BMEI.2016.7852852.
3. P. Ahmadi and M. Joneidi, "A new method for voice activity detection based on sparse representation," 2014 7th International Congress on Image and Signal Processing, Dalian, 2014, pp. 878-882, doi: 10.1109/CISP.2014.7003901.
4. T. Izawa, "Early days of VAD method," 2016 21st OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS), Niigata, 2016, pp. 1-3.
5. R. Ahmad, S. P. Raza and H. Malik, "Unsupervised multimodal VAD using sequential hierarchy," 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Singapore, 2013, pp. 174-177, doi: 10.1109/CIDM.2013.6597233.
6. H. Sahli, L. Tlig, A. Zaafouri and M. Sayadi, "A comparative study applied to dynamic textures segmentation," 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, 2016, pp. 217-222.
7. M. Parada and I. Sanches, "Visual Voice Activity Detection Based on Motion Vectors of MPEG Encoded Video," 2017 European Modelling Symposium (EMS), Manchester, 2017, pp. 89-94, doi: 10.1109/EMS.2017.26.
8. J. Park, Y. G. Jin, S. Hwang and J. W. Shin, "Dual Microphone Voice Activity Detection Exploiting Interchannel Time and Level Differences," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1335-1339, Oct. 2016, doi: 10.1109/LSP.2016.2597360.
9. A. Touazi and M. Debyeche, "A Case Study on Back-End Voice Activity Detection for Distributed Specch Recognition System Using Support Vector Machines," 2014 Tenth In-

ternational Conference on Signal-Image Technology and Internet-Based Systems, Marrakech, 2014, pp. 21-26, doi: 10.1109/SITIS.2014.54.
10. B. Peng and T. Li, "A Probabilistic Measure for Quantitative Evaluation of Image Segmentation," in IEEE Signal Processing Letters, vol. 20, no. 7, pp. 689-692, July 2013.
11. O. Tymchenko, B. Havrysh, O. Khamula, B. Kovalskyi, S. Vasiuta and I. Lyakh, "Methods of Converting Weight Sequences in Digital Subtraction Filtration," 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2019, pp. 32-36.
12. V. A. Volchenkov and V. V. Vityazev, "Development and testing of the voice activity detector based on use of special pilot signal," 2016 5th Mediterranean Conference on Embedded Computing (MECO), Bar, 2016, pp. 108-111.
13. A. Sehgal, F. Saki and N. Kehtarnavaz, "Real-time implementation of voice activity detector on ARM embedded processor of smartphones," 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, 2017, pp. 1285-1290.
14. S. Jelil, R. K. Das, S. R. M. Prasanna and R. Sinha, "Role of voice activity detection methods for the speakers in the wild challenge," 2017 Twenty-third National Conference on Communications (NCC), Chennai, 2017, pp. 1-6, doi: 10.1109/NCC.2017.8077146.
15. K. T. Sreekumar, K. K. George, K. Arunraj and C. S. Kumar, "Spectral matching based voice activity detector for improved speaker recognition," 2014 International Conference on Power Signals Control and Computations (EPSCICON), Thrissur, 2014, pp. 1-4.
16. M. Pandharipande, R. Chakraborty, A. Panda and S. K. Kopparapu, "An Unsupervised frame Selection Technique for Robust Emotion Recognition in Noisy Speech," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, 2018, pp. 2055-2059, doi: 10.23919/EUSIPCO.2018.8553202.
17. A. Moldovan, A. Stan and M. Giurgiu, "Improving sentence-level alignment of speech with imperfect transcripts using utterance concatenation and VAD," 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, 2016, pp. 171-174, doi: 10.1109/ICCP.2016.7737141.
18. H. Kanamori, "Fiber and fiber based technology after VAD development," 2016 21st OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS), Niigata, 2016, pp. 1-3.
19. S. Tong, H. Gu and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5695-5699, doi: 10.1109/ICASSP.2016.7472768.
20. J. Song et al., "Research on Digital Hearing Aid Speech Enhancement Algorithm," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 4316-4320, doi: 10.23919/ChiCC.2018.8482732.
21. D. Peleshko, M. Peleshko, N. Kustra and I. Izonin, "Analysis of invariant moments in tasks image processing," 2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana-Svalyava, 2011, pp. 263-264.
22. Z. Fan, Z. Bai, X. Zhang, S. Rahardja and J. Chen, "AUC Optimization for Deep Learning Based Voice Activity Detection," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6760-6764, doi: 10.1109/ICASSP.2019.8682803.