VAD in Speech Coder Based on Wavelet Transform

Oleksandr Tymchenko [1,2], Bohdana Havrysh [3], Orest Khamula[2] and Bohdan Kovalskyi[2]

[1] *University of Warmia and Mazury, Ochapowskiego str,2, Olsztyn, 10-719, Poland*
[2] *Ukrainian Academy of Printing, Pidholosko st., 19, Lviv, 79020, Ukraine*
[3] *Lviv Polytechnic National University, Stepana Bandery Street, 12, Lviv, 79000, Ukraine*

Abstract

One of the methods of reducing the data flow in packet transmission of speech is to exclude those speech packets that do not carry enough information. This leads to the use of the systems VoIP (Voice over IP) or voice transmission technologies by IP (Internet Protocol – IP) a new type of voice signal encoders – with a variable encoding speed. The functions of detecting packets with a small voice load are carried by nodes with the standard name Voice Activity Detector (VAD). The main difficulty in the synthesis of VAD for VoIP speech signal encoders is the correct detection of speech pauses against a background of sufficiently intense acoustic noise (office, street, car noise, etc.). However, the use of VAD makes it possible to significantly save the bandwidth of a distributed network.

Keywords 1

Speech signal, coder, Voice Activity Detector, Voice over IP

1. Introduction

The use of VAD provides a possible to pre-process the speech signal before transferring it to the encoder. In the first approximation, the following types of signal fragments can be selected: vocalized, non-vocalized, transitional and pauses. When transmitting speech in digital form, that is, in the form of a sequence of numbers, each type of signal with the same duration and the same quality requires a different number of bits for encoding and transmission. Therefore, the transmission speed of different types of signals can be also different. An important conclusion follows from this: the transmission of speech data in each direction of a duplex channel should be considered as the transmission of asynchronous logically independent fragments of digital sequences (transactions) with block (datagram) synchronization within a transaction filled with blocks of different lengths.

2. Overview of existing VAD solutions

VAD technology is used in conbination with a large number of language codecs.

1. The simplest example of the VAD mechanism is illustrated in fig. 1 [1]. The speech signal enters the input of the comparison device, in which its amplitude is measured and compared with a given threshold value at some selected time interval.

When the amplitude of the input signal exceeds the specified threshold (Fig. 1), the signal enters the codec input and encodes according to a certain algorithm (interval T_2–T_3). If the amplitude of the input signal is lower than the threshold value (for example, before the interval T_1–T_2), then at the moment the T_1 transfer only service information (much smaller size) about the beginning of the pause, and at the

MoMLeT+DS 2022: 4th International Workshop on Modern Machine Learning Technologies and Data Science, November, 25-26, 2022, Leiden-Lviv, The Netherlands-Ukraine
EMAIL: olexandr.tymchenko@uwm.edu.pl (O. Tymchenko); dana.havrysh@gmail.com (B. Havrysh); khamula@gmail.com (O. Khamula); bkovalskyi@ukr.net (B. Kovalskyi)
ORCID: 0000-0001-6315-9375 (O. Tymchenko); 0000-0003-3213-9747 (B. Havrysh); 0000-0003-0926-9156 (O. Khamula); 0000-0001-9088-1144 (B. Kovalskyi)

moment T_2 about its end. On the receiving side, during the pause, to improve the subjective perception of speech, a comfortable noise can be submit into the phone [2, 7].

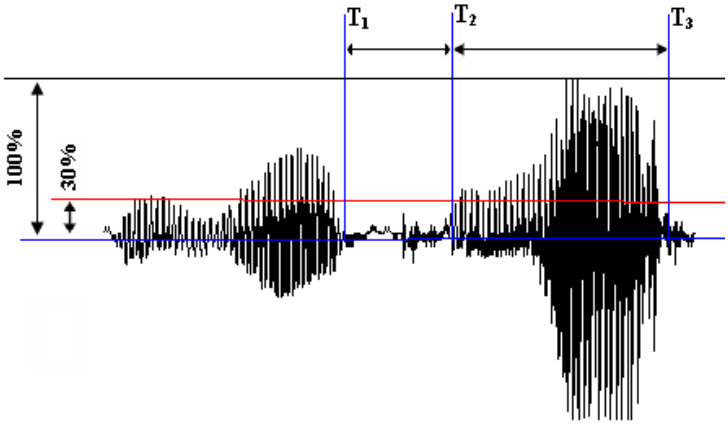

Figure 1: Principle of operation of VAD by level

2. The considered algorithm has been improved in [2, 3], when it is applied to each of the speech segments which formed when the speech signal is broken down into quasi-stationary T_{fr} fragments. Then the set threshold discards (resets) part of the counts of the speech fragment that are smaller than it. After that, after counting the number of rejected counts, an additional threshold is set, determined by the quantitative ratio between existing counts in the speech fragment, to counts that are smaller than the selected threshold and subordinate to rejection. Based on the results of the analysis, a decision is made about the possibility of discarding the entire package, which significantly increases the effectiveness of the VAD as a whole.

The algorithm of such a VAD will be presented as follows. For each speech fragment, the following is calculated (1):

$$m = \sum_{i=1}^{k} \left(x_i < x_{tv} \right) \tag{1}$$

where k – the total number of readings in the fragment; x_{tv} – threshold value

The limit number of readings in a speech fragment is set – $m_{tv} < k$, less than which is considered that the fragment does not contain speech load. If $m < m_{tv}$, then a decision is made to discard (reset) this speech fragment.

3. There is a VAD method based on the power level calculated in each speech segment P_{fr} [3]. This algorithm is more complicated in comparison to the previous one, as it requires the calculation of the power of the running fragment and the determination of the ratio (2):

$$q = \frac{P_{fr}}{P_{\max}} = \frac{\sum_i x_i^2}{\sigma_x^2} \tag{2}$$

If the value is less than the specified threshold $q < q_{tv}$, then, as in the previous case, this will determine the need to discard the packet.

1. The GSM standard adopts the VAD scheme with processing in the frequency domain [4, 5]. The structural diagram of such a VAD is shown in Fig. 2. Her work is based on the difference between the spectral characteristics of speech and noise. Background noise is assumed to be stationary over a relatively long period of time, and its spectrum slowly changes over time. VAD determines the spectral deviations of the input action from the background noise spectrum.

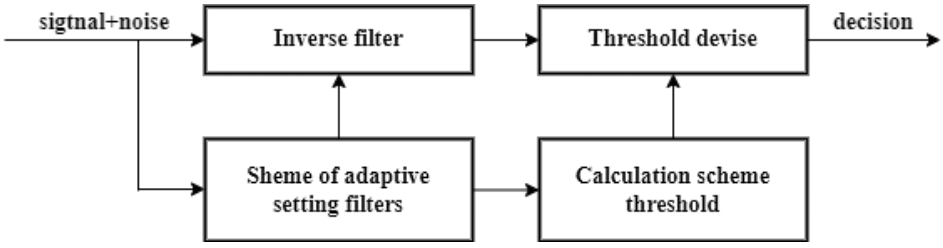

Figure 2: Structural diagram of the VAD according to the spectral characteristics of the noise

This operation is carried out by an inverse filter, the coefficients of which are set in relation to the action at the input of the circuit for predicting only background noise. If there is a speech signal and noise at the input, the inverse filter suppresses the noise components and reduces its intensity. The energy of the signal + noise at the output of the inverse filter is compared with the threshold. This threshold is higher than the energy level of the noise signal only. Exceeding the threshold is considered for presence of a signal. The coefficients of the inverse filter and the threshold level change over time depending on the current value of the noise level when only noise is applied to the input. Since these parameters (coefficients and threshold) are used by the VAD detector for speech detection, the VAD itself cannot make a decision at this stage of the analysis, since the threshold can change [6-9].

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

VAD with processing in the spectral domain is successfully combined with the speech RPE/LTP-LPC codec, since the envelope of the input action spectrum, necessary for the operation of the secondary VAD, is already determined in the process of LPC analysis. An obvious disadvantage of this scheme is the "relatively long period of time" during which a decision is made about the presence of noise or a signal.

3. Basic requirements for the VAD system

The VAD detector must be sensitive and fast to avoid loss of word onsets when transitioning from a pause to an active speech fragment, at the same time the VAD detector must not be triggered by background noise.

The work of the VAD is to estimate the value of the parameter of the input signal (for example, level, power or energy) and if it exceeds a certain threshold, activate the transmission of the packet. This slightly increases the delay when processing the speech signal in the codec, but it can be minimized by using encoders that work with blocks of counts, for example, based on transformations.

In the coder analyzer with the language coding rate C (bit/c), the signal is segmented into separate fragments (as a rule, quasi-stationary sections) with duration T_{fr} from 2 to 50 ms, and for each input block consisting of N counts, is created accordingly an information frame of size (3):

$$V_k.(bit) = T_{fr}.\cdot \mathrm{C}_k, \quad (3)$$

Regardless of the details of the implementation, the main determinant in the evaluation of the encoder is the high quality of speech reproduction at a low speed of the output digital stream C_k with minimal requirements for the resources of the digital processing processor and minimal delay.

3.1. Determination of the main parameters of the telephone signal for the operation of the VAD system

A telephone conversation between two consumers is usually accompanied by a large number of emotional pauses, and taking into account the delays experienced by speech packets in VoIP networks (more than 50 *ms*), pauses are added to natural speech to wait for an answer.

In telephone networks, for the transmission of speech signals with high quality (with satisfactory naturalness and intelligibility of syllables – 90% and phrases – 99%), you can limit yourself to the frequency band of 0.3...3.4 kHz [2, 10-12].

For this study, it can be assumed that the distribution of instantaneous values of speech signals corresponds to the exponential law [3] (4):

$$F(x)=\begin{cases} e^{ax}/2, x\geq 0; \\ 1-e^{ax}, x<0, \end{cases} \quad (4)$$

where $F(x)$ – the probability of instantaneous signal values; X_{tf} – the effective (root mean square) value of the signal $x(t)$ (5).

$$X_{tf}=\sqrt{P_{tf}\cdot 1\ Ohm}=\sqrt{\frac{1}{T_c}\int_0^{T_c}x^2(t)dt} \quad (5)$$

where P_{tf} – the signal power averaged over the observation time $T_c = T_{fr}$.

For a telephone signal (TF), it is possible to accept the exponential law of the distribution of instantaneous values of speech signals with the parameter $a=\sqrt[3]{2}/X_{tf}$ then, for the value of the limit voltage X_{lim}, the probability F of exceeding which is $\varepsilon < 10^{-3}$, it is possible to write (6):

$$e^{aX_{tf}}=2\varepsilon \text{ where } X_{lim}\approx 5X_{tf}. \quad (6)$$

In this case, the value of the signal peak factor is equal to (7):

$$Q_{tf}=10\ lg\left(\frac{P_{tf.max}}{P_{tf}}\right)=20\ lg\left(\frac{X_{lim}}{X_{tf}}\right)\approx 14\ dBW \quad (7)$$

Relation (8):

$$Y_{tf}=10\ lg\left(\frac{P_{tf}}{P_{meas}}\right)\ dBW \quad (8)$$

is the dynamic level (volume) of the TF signal. In this expression P_{meas} is the power of the measured signal at the point of the tract where research is conducted. According to the ITU-T recommendations, volumes are measured with a special device (volume meter), which provides a quadratic law of summation of oscillations of different frequencies, has a logarithmic scale (in decibels) and a time constant (integration time) of $T_c = 200\ mc$. Statistical studies have established that the distribution of volumes obeys the Gaussian law with an average value

$Y_{tf.aver}=-12.7\ dBmW$ and the mean square deviation $\sigma_Y = 4.3\ dB$.

$$w_{(Y)}=\frac{1}{\sqrt{2\pi\sigma_Y}}e^{-\frac{\left(Y_{tf}-Y_{tf.aver}\right)^2}{2\sigma_Y^2}} \quad (9)$$

where $w_{(Y)}$ – the volume distribution density; σ_Y – the root mean square deviation.

The corresponding average power level in the voice frequency circuit can be found using the formula (10):

$$P_{tf.aver}=Y_{tf.aver}+\frac{lg10}{20}\sigma_Y^2=-12.7+0.115\cdot 4.3^2=-10.57\ dBmW \quad (10)$$

Then $P_{tf.aver}=10^{0.1(-10.57)}=88\ \mu W$ – the average power of the TF signal in the voice frequency circuit without taking pauses into account.

The influence of pauses is taken into account using the activity coefficient K_a of the signal source. It is equal to the ratio of the time during which the signal level at its output exceeds the set threshold value (usually –40 dBmW) to the total conversation time. Statistics of the TF signal gives $K_a \geq 0.25$. Then the average power of the TF signal, including pauses (11):

$$P_{tf.aver.n}\approx K_aP_{tf.aver}+10,$$
$$P_{tf.aver.n}=32\mu W\ (-15\ dBmW) \quad (11)$$

here the second term of the right-hand side, equal to 10 μW, is introduced according to the ITU-T recommendations, as a correction for the increased power of the signals accompanying the TF conversation (staff conversations and service signals transmitted on the same channel).

Taking into account expression (6), it is possible to determine the maximum level of $P_{tf.max}$. corresponding to the maximum power $P_{tf.max}$ and the limit voltage X_{lim} (12):

$$P_{tf.max}=P_{tf.aver}+Q_{tf}-10.57+14=3.43\ dBmW\ . \quad (12)$$

For signals transmitted via digital channels, $P_{tf.max} = +3\ dBmW$ is usually accepted, and for signals transmitted using analog transmission systems $+3\ dBmW$. In the latter case, the maximum power of $P_{tf.max}$ will be equal to 2220 μW.

The minimum volume is considered to be the volume whose smaller values appear with a probability of $\varepsilon<10^{-3}$ (13):

$$Y_{tf.min} = Y_{tf.aver} - 3.09\sigma_Y. \quad (13)$$

Then the level of $P_{tf.min}$, which corresponds to the minimum signal, will be lower than $Y_{tf.min}$ by the value of the peak factor. Thus, the dynamic range of the $D_{c.tf}$ signal, taking into account formulas (10) and (12), will be (13):

$$D_{c.tf} = P_{tf.max} - P_{tf.min} = 2Q_{tf} + 3.09\sigma_Y + 0.115\sigma_Y^2 = 40\ dB \quad (13)$$

It was experimentally established [3, 13, 15-17] that the quality of reception of the TF signal is still acceptable at an average interference power of 178000 pW. Therefore, the required immunity of the telephone signal should be (14):

$$A_{im.tf} \geq 10\lg\left(88 \cdot \frac{10^{-6}}{178000} \cdot 10^{-12}\right) \approx 27\ dB \quad (14)$$

The dynamic range of the TF signal, calculated as the ratio of the maximum power to the average power of the permissible fluctuation interference, is equal to (15):

$$D_{c.tf} = 10\lg\left(2220 \cdot \frac{10^{-6}}{178000} \cdot 10^{-12}\right) \approx 41dB, \quad (15)$$

which slightly differs from the value found by formula (10).

When evaluating the potential information volume, it is necessary to take into account the activity coefficient of the signal source. Then for the speech signal (16):

$$V_{tf.max} = K_a F_B log_2\left(1 + 10^{0.1A_{im.tf}}\right) = 7.6\ kbps \quad (16)$$

Here, $F_B = 3.4\ kHz$ is the upper frequency of the tone frequency channel.

3.2. A proposed mechanism of VAD

In the VAD structure (Fig. 2), the noise level at the output of the encoder is equalized and its threshold is set, adapting the encoder in relation to the value of the input speech signal. Analogous to this solution, noise equalization can be replaced by maintaining the input signal level within certain limits, while simultaneously fixing the noise level. That is, such a scheme becomes adaptive to the signal level. The implementation of such a principle is simple, both in analog and digital form. Its structure corresponds to the circuit with automatic gain control (ARP), which means it has the ability to adapt almost instantly to the level of the input signal.

The very implementation of the proposed principle of VAD operation is faster and simpler than in the GSM system. The integrator present in the circuit must have a time constant $\tau = 0.5...1s = T_{RC}$ and the condition $T_{RC} > T_{tf}$ must be fulfilled. The noise level will be maintained at the protection level of the tone frequency channel $A_{im.tf} \approx 27dB$. At the same time, it is not necessary, as in GSM, to make a delay (for five packets) to start transmitting a sign of comfort noise.

4. Research and modeling of the proposed voice activity detector

Considering the properties of the speech signal, the need to use VAD is obvious when creating and researching models of speech signal encoders [4, 14, 18] for VoIP. In the process of modeling the coding method based on the wavelet transformation, a processing block was introduced, which implements the speech activity detector function highlighted in Fig. 3 (a). The VAD operation algorithm makes it possible to reset speech count packets if they contain counts smaller than the threshold set by the VAD block.

4.1. VAD modeling by signal level

1. Beforehand a calculation is made for X_{max} of the maximum signal count in all segments of the speech fragment, that can be processed (by default $X_{max} = 1$);
2. x_{lev} is established. (expressed in the number of levels k_{lev} = 1, 2, 4, 8, 16, 32) in relation to X_{max} (17)

$$x_{lev} = \frac{k_{lev}}{k_{max}} X_{max}, \tag{17}$$

where $k_{max} = 127$ – the maximum number of levels;

a comparison of all counts in the segments x_{lev} and determined their numbers $m = \sum_{i \in k}(x_i < x_{lev})$, further, according to the set threshold for the number of readings that can be discarded in the segment (for example, 80%), which is smaller than the percentage, if $m < K \cdot 0.8k_{x_{lev}}$, ($K$=160 counts are selected in the research), then it is decided that all readings of the segment are equal zero;

3. after processing one segment, we move to the next (up to 2) until we check the entire file.

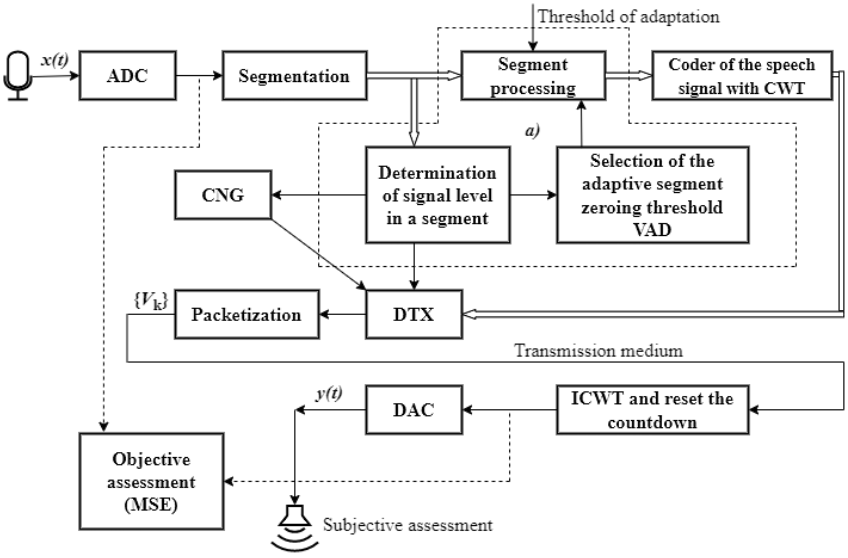

Figure 3: The scheme of evaluating the quality of the speech signal VoIP with WT highlighted (a) VAD structure

4.2. Modeling VAD by power

The algorithm model implements the following sequence of operations:

1. estimation of the maximum signal power in the speech fragment that can be processed by the VAD detector;
2. setting the threshold k_P= 0.2, 0.3, 0.5, 0.75, 1, 2 in % of P_{max};
3. performing a comparison of the signal powers in each segment of P_i with the selected threshold – $P_i < k_P P_{max}$ and if the condition is met, then this packet is discarded;
4. the operations of points 2, 3 are repeated for all segments of the speech fragment.

The speech segments processed according to the proposed VAD algorithms were transfer to the coder that uses the wavelet transformation, after which the reverse decoding and assembly of the segments was carried out. To assess the quality of speech signal reproduction, the objective assessment proposed in [19, 20] was used by comparing input and output speech readings and determining the root mean square deviation (SKD) between them. It is also possible to listen to the source file, that is, to subjectively evaluate the received speech signal by ear.

The results of modeling and research are presented in fig. 4-8. Shown: fig. 4 – signal form; Fig. 5 – (SKD) of the output signal; Fig. 6 – compression coefficient K_{st}; Fig. 7 – relevant parameters for VAD

by power; Fig. 8 – comparison of VAD methods by level and power. (MSE) values were noted and K_{st} values were achieved with reasonably good legibility.

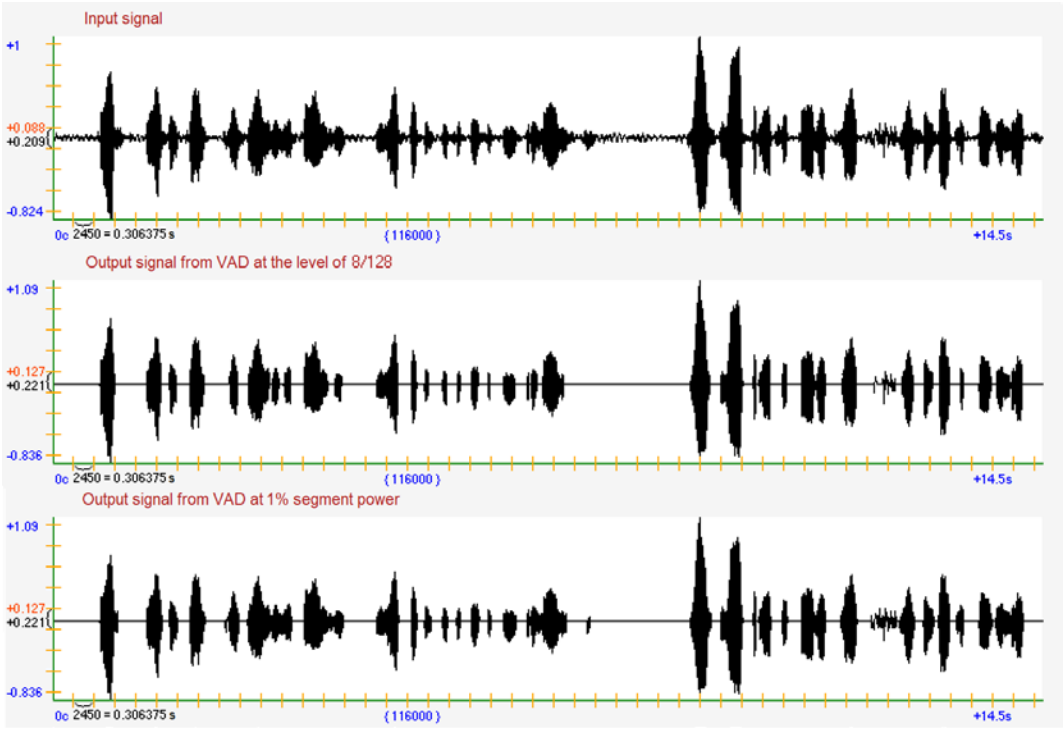

Figure 4: VAD performance study

a) input signal;

b) output signal from VAD at the level of 1/16;

c) output signal from VAD at 1% segment power

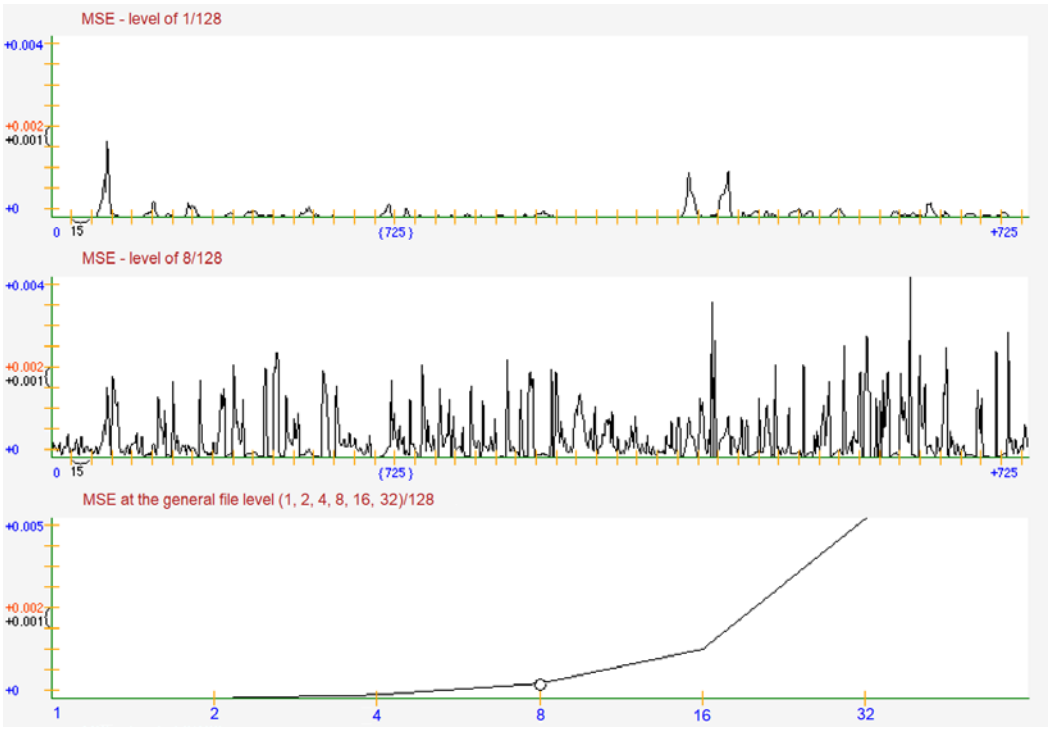

Figure 5: Study of the operation of VAD MSE of the output signal

a) VAD at the level of 1/128;

b) VAD at the level of 1/16;

c) VAD at the general file level

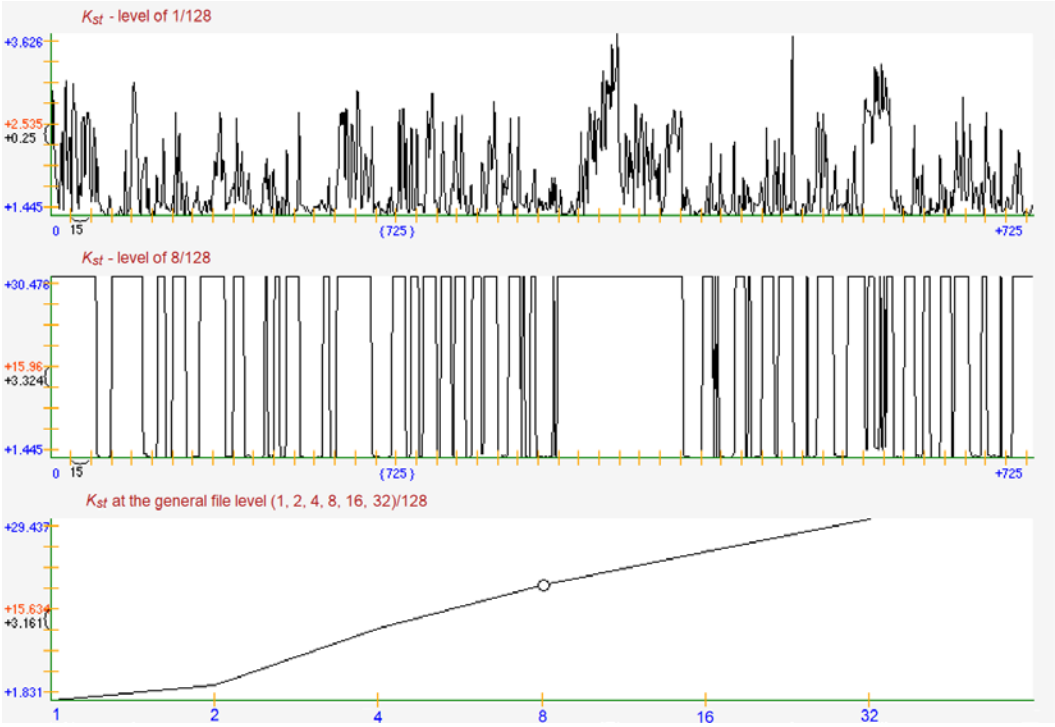

Figure 6: Study of the work of the VAD compression ratio

a) VAD at the level of 1/128;

b) VAD at the level of 1/16;

c) VAD at the general file level

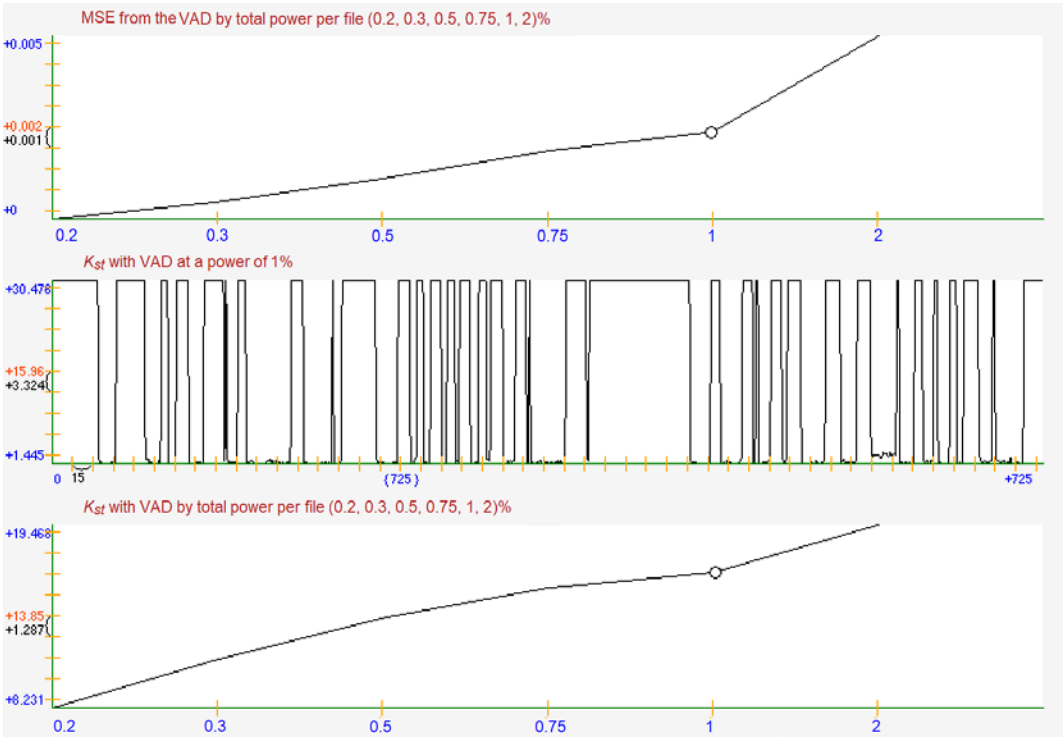

Figure 7: VAD performance study

a) MSE of the output signal from the VAD by power common to the file;

b) compression ratio with VAD at a power of 1%;

c) compression ratio with VAD by total power per file

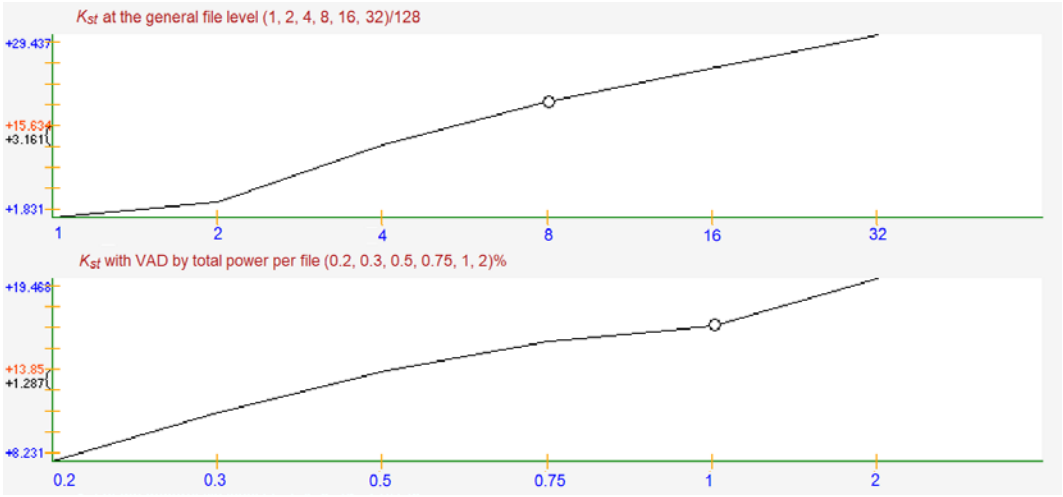

Figure 8: Comparison of VAD methods by compression ratio
a) VAD by level;
b) VAD by power

5. Conclusion

The proposed VAD operation algorithm is highly efficient and provides a sufficiently high compression ratio. When evaluating the quality of VAD work according to the SKD criterion and when listening to processed speech samples, it is possible to indicate the maximum value of rejected quantization levels and the value of the adaptive power threshold at which the number of zeroed segments practically does not affect the quality of the restored speech signal. Therefore, during the conducted research, slight deviations in the quality of speech signal reproduction are observed when discarding up to 8 quantization levels out of 128, or when a power threshold of 0.3% of the maximum signal level in the segment is set. Accordingly, the achieved compression ratio is 19.35 and 15.6 (bit rate 3.3 and 4.1 kbps).

6. Acknowledgments

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

7. References

[1] G. Jung, N. Cho, H. Kim and H. Cho, "DNN-GRU Multiple Layers for VAD in PC Game Cafe," 2018 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), 2018, pp. 206-212, doi: 10.1109/ICCE-ASIA.2018.8552099.

[2] 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), 2016, pp. 5695-5699, doi: 10.1109/ICASSP.2016.7472768.

[3] V. Kovtun, O. Kovtun, and A. Semenov, “Entropy-Argumentative Concept of Computational Phonetic Analysis of Speech Taking into Account Dialect and Individuality of Phonation,” Entropy, vol. 24, no. 7. MDPI AG, p. 1006, Jul. 20, 2022. doi: 10.3390/e24071006.

[4] P. Mayorga, G. Chavez, C. Druzgalski, V. Zeljkovic and J. Valdez, "Lung sounds focused events classification and segmentation with VAD-GMM," 2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), 2017, pp. 1-5, doi: 10.1109/GMEPE-PAHCE.2017.7972086.

[5] M. Nazarkevych, Y. Voznyi, V. Hrytsyk, I. Klyujnyk, B. Havrysh and N. Lotoshynska, "Identification of Biometric Images by Machine Learning," 2021 IEEE 12th International

Conference on Electronics and Information Technologies (ELIT), 2021, pp. 95-98, doi: 10.1109/ELIT53502.2021.9501064.
[6] J. Song et al., "Research on Digital Hearing Aid Speech Enhancement Algorithm," 2018 37th Chinese Control Conference (CCC), 2018, pp. 4316-4320, doi: 10.23919/ChiCC.2018.8482732.
[7] M. Shovheniuk, B. Kovalskiy, M. Semeniv, V. Semeniv, N. Zanko. Information Technology of Digital Images Processing with Saving of Material Resources. Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume I: Main Conference. Kherson, Ukraine, June 12-15, 2019. pp. 414-419. http://ceur-ws.org/Vol-2387/20190414.pdf.
[8] 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), 2019, pp. 32-36, doi: 10.1109/STC-CSIT.2019.8929750.
[9] S. A. Wibowo and K. Usman, "Voice activity detection G729B improvement technique using K-Nearest Neighbor method," 2010 International Conference on Distributed Frameworks for Multimedia Applications, 2010, pp. 1-5.
[10] W. Y. Yuan, Y. Zhou, Z. Y. Huang and H. Q. Liu, "A VAD-based switch fast LMS/Newton algorithm for acoustic echo cancellation," 2015 IEEE International Conference on Digital Signal Processing (DSP), 2015, pp. 967-970, doi: 10.1109/ICDSP.2015.7252021.
[11] L. K. Hamaidi, M. Muma and A. M. Zoubir, "Multi-speaker voice activity detection by an improved multiplicative non-negative independent component analysis with sparseness constraints," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 4611-4615, doi: 10.1109/ICASSP.2017.7953030.
[12] M. Shahid, C. Beyan and V. Murino, "Voice Activity Detection by Upper Body Motion Analysis and Unsupervised Domain Adaptation," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 1260-1269, doi: 10.1109/ICCVW.2019.00159.
[13] B. Durnyak, O. Tymchenko, O. Tymchenko and B. Havrysh, "Applying the Neuronetchic Methodology to Text Images for Their Recognition," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018, pp. 584-589, doi: 10.1109/DSMP.2018.8478482.
[14] S. Hizlisoy and Z. Tufekci, "Noise robust speech recognition using parallel model compensation and voice activity detection methods," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818517.
[15] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis and I. Pitas, "Visual Voice Activity Detection in the Wild," in IEEE Transactions on Multimedia, vol. 18, no. 6, pp. 967-977, June 2016, doi: 10.1109/TMM.2016.2535357.
[16] M. Wake, M. Togami, K. Yoshii and T. Kawahara, "Integration of Semi-Blind Speech Source Separation and Voice Activity Detection for Flexible Spoken Dialogue," 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 775-780.
[17] M. Yang et al., "Nanowatt Acoustic Inference Sensing Exploiting Nonlinear Analog Feature Extraction," in IEEE Journal of Solid-State Circuits, vol. 56, no. 10, pp. 3123-3133, Oct. 2021, doi: 10.1109/JSSC.2021.3076344.
[18] Y. Hou, Y. Deng, B. Zhu, Z. Ma and D. Botteldooren, "Rule-Embedded Network for Audio-Visual Voice Activity Detection in Live Musical Video Streams," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 4165-4169, doi: 10.1109/ICASSP39728.2021.9413418.
[19] A. Hassani, A. Bertrand and M. Moonen, "Real-time distributed speech enhancement with two collaborating microphone arrays," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 6586-6587, doi: 10.1109/ICASSP.2017.8005295.
[20] O. Tymchenko, B. Havrysh, O. O. Tymchenko, O. Khamula, B. Kovalskyi and K. Havrysh, "Person Voice Recognition Methods," 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), 2020, pp. 287-290, doi: 10.1109/DSMP47368.2020.9204023.