ARCHIVES OF ACOUSTICS
19, 1, 5–35 (1994)

SPEECH SIGNAL TRANSMISSION RATE COMPRESSION USING OF THE TIME PARAMETERS CODING METHOD

CZ. BASZTURA

Acoustic signals analysis and processing division
Institute Telecommunications and Acoustics
of the Technical University of Wrocław
(50-370 Wrocław, ul. Janiszewskiego 7/9)

A new method of speech signal encoding and decoding is the subject of the presented research project. This method may find applications in the field of telecommunications (for telephone or radio transmission) as well as in the field of speech synthesis. The concept of the method is based on the extraction and transmission of such time parameters as the intervals between the subsequent speech signal zero-crossings and the amplitude of the signal in these intervals as well as on the subsequent speech signal's reconstruction based on the given parameters and knowledge derived from the analysis of speech. The system is composed of two main processes: the extraction and encoding of the transmission parameters and the original speech signal reconstruction (resynthesis). The method makes it possible ot decrease the transmission rate about 10 times, as compared to the original speech signal. The results prove that the synthesized speech signal quality, when the new method is used, may be better than the one obtained by the use of other vocoder methods. The method's low cost and a relatively simple hardware system developed for the parameters' extraction and for the reconstruction of the speech signal are the most important advantages of the described method.

1. Introduction

1.1. Presentation

The switching to digital processing, encoding and transmission circuits and systems began as early as in the beginning of the sixties when the first system of the encoded information transmission was presented. The information was passed with the decreased error rate as well as the transmission rate.

The interest and demand of efficient coding methods and speech signal transmission rate compression as well as the transmission quality estimation grew along with the progress and development of the new speech signal coding for teletransmission purposes [1, 2, 3, 5, 6, 7, 11, 12, 13, 18, 29]. As compared to the analog information transmission, the digital method has a number of

advantages including the possibilities of detecting and correcting errors and also better distortion protection.

Digital speech signal transmission systems are divided into two categories:

1) Waveform encoding systems,
2) Particular speech signal parameters encoding systems.

The latter are called vocoders or source coders [23].

1.2. The concept of the speech signal transmission rate compression method based on time parameters coding

The two categories of coding described in Section 1.1. were applied in digital teletransmission systemes. Both methods are included in the international recommendation CCITT, the pulsecode modulation (PCM) at the bit rate of 64 Kbit/s as well as the adapted differential PCM at the rate of 32 Kbit/s.

Intensive research is being carried out in order to develop relatively little complicated vocoder systems which have a transmission rate less than 16 Kbit/s. At the same time the aim is to preserve the processed signal's quality as similar to the one achieved by using PCM or ADPCM methods. The solution to this problem is being searched in the parametric encoding method. The parametric encoding method bound with the speech signal information volume (transmission rate) compression consists in isolating from speech signal certain parameters that describe its amplitude variations and the component frequencies of the spectrum as well as the transmission of these parameters actual values to the receiver by a telecommunication channel instead of the original speech signal. The parameters isolated from the speech signal usually form slow changing courses; hence for their transmission a channel of a much lower bit rate may be used than the one which transmits the original signal. This is the method that is used in encoding, transmission and decoding in vocoder technology [1, 2, 8, 9, 25].

The known methods of parametric coding, i.e., speech signal transmission rate compression based on harmonic spectral-band formant and linear predicted coding methods, make it possible to achieve a speech signal transmission rate of the range from 4.8 Kbit/s to 1 Kbit/s. The concept of a new speech signal transmission rate compression is based on the zero crossing analysis (ZCA). There are two main premises for this concept:

1) Theoretical, described, among others, in the publication by Logan [15] and Basztura [4] as well as in publications by other authors concerning the possibility of reconstructing (with a certain inaccuracy) a limited bandwidth signal on the basis of knowledge about zero crossings,
2) Practical, resulting from the fact that the parameters based on ZCA and describing linguistic differences of linguistic units as well as their particular features in automatic voice and speech recognition processs are the parameters of similar discrimination power to other — more complex parameters as spectral and predictive ones [4, 10].

The substance of the time vocoder's concept, i.e., speech signal transmission rate compression based on time parameters encoding, is the asynchronical isolation of such basic parameters as lengths of subsequent intervals between the zero-crossings I_n and amplitudes of courses in these intervals A_n form the digital signal. These two parameters are transmitted by a telecommunication link to a receiver where the signal is resynthesized (reconstructed) basing on the received sequences $\{I_n, A_n\}$. They may be additionally quantized, what makes it possible to diminish the information volume. These data are supplemented by the information obtained during preliminary research and analyses residing in the resynthesis circuit's memory. This allows to recreate the input (to the receiver) speech signal U_n as the signal U'_n corresponding very well with the input signal. The method efficiency estimation criterion is, above all, the relation between signal transmission rate compression rate and the output signal quality estimation U'_n. The resynthesized signal estimation comprises a subjective estimation (including intelligibility, the transmission of individual features and the quality of sound) and objective estimation [2, 11, 12, 13, 28].

The planned research had the following purposes:

1) the statistical (for stationary segments of speech) and deterministic (for each of the Polish speech phonemes) data elaboration regarding the values of: time intervals resolution between zero-crossings of the speech signal I_n, the values and probabilities of occurrence of A_n amplitudes in these intervals,
2) the determination of the possibility of scalar or vector quantization for $\{A_n, I_n\}$ allowing further transmission rate compression,
3) the design of the digital simulation model of a vocoder,
4) carrying out research concerning subjective distinctness and intelligibility in order to set a selection of optimal parameters and the estimation of the vocoder's quality,
5) carrying out tests on objective compatibilities of the input signal U_n and the resynthesized signal U'_n,
6) determining the vocoder's primary parameters and presenting the concept of the method's implementation possibilities.

2. Methods and systems of digital speech signal encoding

2.1. Presentation

All wave from encoding devices and digital vocoders have one feature in common, i.e., analog-to-digital and digital-to-analog processing application. It has to be admitted that all imperfections introduced by these converters may reduce significantly the benefits of digital speech signal processing. The importance of selecting the converters' properties is motivated by a few factors:

- too high quality of the converter increases costs and may result in useless engaging of the computing capacity during further steps of signal processing,

- the measured decrease (increase) of quality (objective) might to correspond with the decrease (increase) of the individually perceived quality.

2.2. *Waveform encoding*

DM and PCM are the most common methods of waveform encoding. Pulse-code modulation is based on the fact that each sample is transmitted by means of the code created form the group of impulses (Fig. 1.). The advantage of this encoding method is a high distortion resistance rate, its disadvantage, though, is a relatively big information volume of the encoded form of the speech signal. The sample frequency $f_{pr} \geqslant 2\, f_g$, where f_g — maximum transmission frequency. Usually $f_g = 4$ kHz is assumed according to the telephone transmission bandwidth.

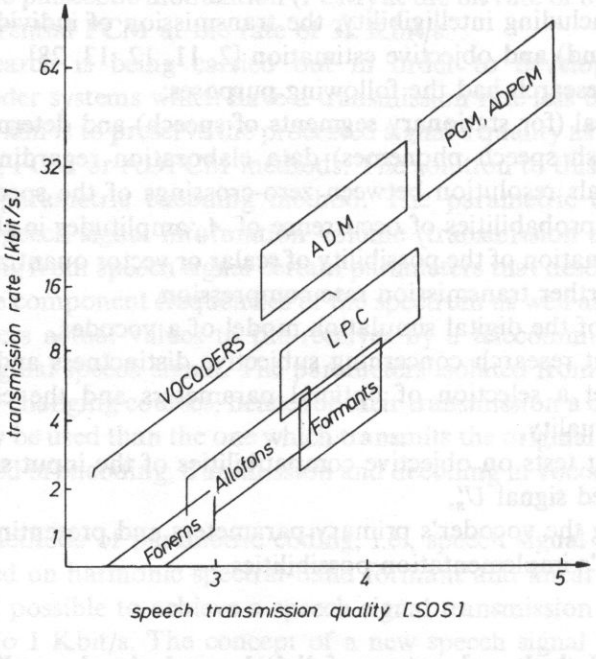

Fig. 1. The different parameters of speech coding methods.

The Delta Modulation (DM) is a way of encoding where the difference of levels between two samples is expressed digitally. The main advantage of this way of processing is an extremely simple circuit realisation but PCM converters do not have to have logical ciruits generating subsequent approximations, amplifying or sampled-data-memorizing filters, introductory amplifiers and other high precision elements [20].

In differential pulse-code modulation (DPCM) the signal increments during Δt intervals between samples as well as their signs are given. In adapted delta modulation the increment's value increases or decreases depending on the signal's changes.

Table 1. The parameters of the selected digital speech signal encoding systems

Speech signal encoding method	Transmission rate [Kbit/.s]	Signal/noise ratio [dB]
Waveform encoding		
Linear pulse-code modulation (PCM)	128	43
Logarithmic pulse-code modulation	56	29
Linear-logarithmic pulse-code modulation	32	29
Linear Delta Modulation (DM)	56	21
Adapted Delta modulation (ADM)	24–16	18
Parametric encoding		
Encoding method	Transmission rate [Kbit/s]	word intelligibility [%]
Spectral-band encoding	4.8–4.2	92
Orthogonal (harmonic) encoding	2.4	86
formant encoding	1.2	80
predicted encoding	4.8–1.2	88

2.3. Parametric coding — vocoders

Parametric encoding is bound with speech signal information companding. This consists in isolating from the speech signal certain parameters describing its amplitude changes as well as the spectrum component frequencies and the transmission of their instantaneous values (in an analog or digital way) by a telecommunication channel instead of a proper speech signal. The parameters isolated from the signal form slow-changing courses; therefore, for their transmission a channel of information volume less than for the proper speech signal transmission may be used. The speech signal at the receiver is subject to reconstruction to the shape approximate to the primary one, therefore it may be assumed that its volume was subject to compression and then — to expansion. At present the following methods of speech signal parametric encoding are known: band spectral (or: band-channel) and harmonic encoding, formant predictive encoding and a few mixed methods developed from them like PARCOR [14, 21]. All types of vocoders mentioned above are characterized by the following disadvantages limiting so far their range of application.

These are:

a) the complexity of encoding and decoding procedures which causes the high cost of emitting and receiving devices,
b) usually non-natural quality of sound,
c) lack of transmission of the individual voice features which results in the fact that the collocutor cannot be recognized by his voice.

Figure 1 shows the dependence of the transmission rate and the speech transmission quality expressed in the listeners' scale of five grades (SOS) from the encoding methods [23]. At present the most developed and preferred vocoders are those based on linear predictive coding. As compared to others, they have many advantages, nevertheless they require special complex processors LPC working in the real time. In spite of the differences, all of the parametric coding methods show many common features. For example, all of them are based on the phonetic-acoustic speech signal microstructure analysis and the isolation of those parameters which describe univocally this microstructure.

The time vocoder has the advantages of waveform encoding methods such as PCM (pulse-code modulation), ADPCM (adapted pulse-code modulation), DM (delta modulation) or ADDM (adapted delta modulation) and at the same time it can ensure a signal information volume compression comparable to those of LPC vocoders. On the ground of the time structured information transmission, the individual pronounciation character is preserved, which is an important element in human intercommunication.

3. Transmission rate compression based on the analysis of time parameters

3.1. Presentation

The concept of the new parametric encoding method, outlined in Section 1.2. of this publication can, be illustrated schematically as it is in the Fig. 2 where the block

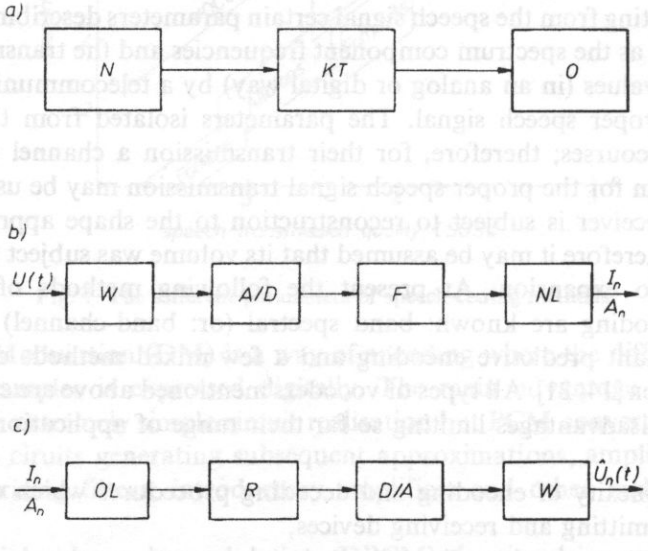

Fig. 2. The block scheme illustrating the conception of the speech signal transmission rate compression method based on coding of the time parameters. *W* — amplifier, *F* — filter, *NL* — line driver, *OL* — line receiver, *R* — resynthesis.

scheme of the vocoder system based on the analysis of time parameters shown. The system encloses the transmitter unit (N) and the receiver (O) unit linked by the digital telecommunication line (KT).

The isolation of the numerical values' sequence of two quantities, i.e., the time intervals' length between the subsequent speech signal's zero-crossings (I_n) and the amplitudes in these intervals (A_n) from the time course is made in the transmitter unit (the analyzer). The pair of these numbers: the interval (I_n) and the amplitude (A_n) are transmitted to the receiver (the synthesizer) where to each interval (I_n) and amplitude (A_n) the course of the actual values is assigned (Fig. 3). The information of the course's shape comes from the signal's statistical and deterministic analyses of Polish speech.

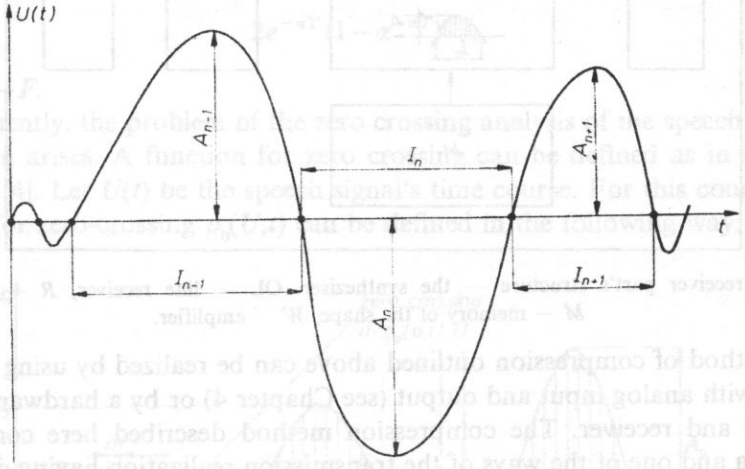

Fig. 3. The basic speech signal's time parameters I_n, A_n.

3.1.1. The transmitter unit — the analyzer

The analyzer (Fig. 4) comprises: an amplifier, analog-to-digital converter with sampling parameters set to $f_{pr} = 10$ kHz and dynamics D (D described on 10 – 12 bits per sample). These are the minimal typical processing parameters A/D for the speech signal as well as for the time intervals extractor, the amplitude and the line transmitter.

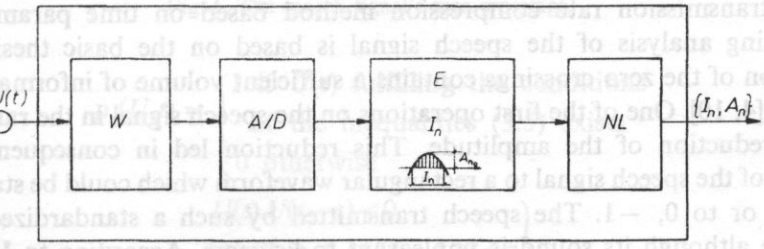

Fig. 4. The transmitter part's structure — the analyzer. W — amplifier, E — extractor, NL — line driver.

3.1.2. *The receiver unit — the synthesizer*

The synthesizer (Fig. 5) comprises: the line receiver, the resynthesizer system, the actual courses shapes' memory (or the processor computing current values of $U_n(t)$ — the rule of synthesis) as well as the D/A converter the parameters of which are compatible with the A/D converter in the analyzer plus the output system.

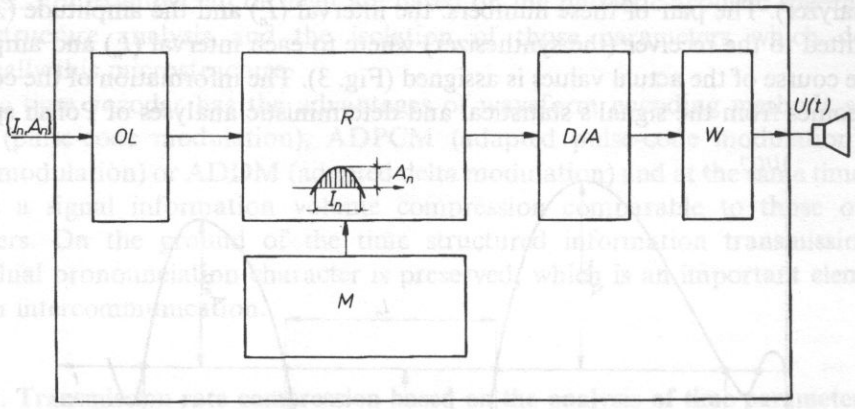

Fig. 5. The receiver part's structure — the syntheziser. OL — line receiver, R — resynthesizer, M — memory of the shape, W — amplifier.

The method of compression outlined above can be realized by using a universal computer with analog input and output (see Chapter 4) or by a hardware system of transmitter and receiver. The compression method described here comprises the general idea and one of the ways of the transmission realization having a significant transmission rate compression. There are some possible modifications to this method that should be examined, e.g., the possibility of transmitting the interval's values I_n alone from the transmitter to the receiver. The A_n values would then be assigned automatically in the receiver on the basis of previously gained data and rules obtained from statistical and deterministic analyses (the correlation $A_n = f(I_n)$ etc.).

3.2. *Zero-crossing analysis*

The premises resulting from zero-crossing analysis form the theoretical base for the new transmission rate compression method based on time parameters. The zero-crossing analysis of the speech signal is based on the basic thesis that the distribution of the zero crossings contains a sufficient volume of information about the signal [4, 10]. One of the first operations on the speech signal in the range of time was the reduction of the amplitude. This reduction led in consequence, to the reduction of the speech signal to a rectangular waveform which could be standardized to $+1$, 0 or to 0, -1. The speech transmitted by such a standardized signal is intelligible although its sound is unpleasant to listeners. According to LOGAN [15], a signal

$$U(t)=z(t)+\cos Ft, \tag{3.1}$$

where $z(t)$ is a bandlimited signal $(-f_g, f)$ and $0<f<F<\infty$ and at the same time it fulfils the conditions of the inequality (3.2), can be reproduced if

$$(-1)^j U(kT/F)>0, \tag{3.2}$$

and

$$|z(t)|<1.$$

The zeros within the time interval $t-T$ and $t+T$ make it possible to reproduce $U(t)$ with a relative error less than

$$2e^{-\alpha T}(1-\alpha^{-\alpha T})^{-2}, \tag{3.3}$$

where $\lambda=f-F$.

Consequently, the problem of the zero crossing analysis of the speech signal and of extraction arises. A function for zero crossing can be defined as in BASZTURA's publication [4]. Let $U(t)$ be the speech signal's time course. For this course (Fig. 6) a function for zero-crossing $\rho_0(U,t)$ can be defined in the following way:

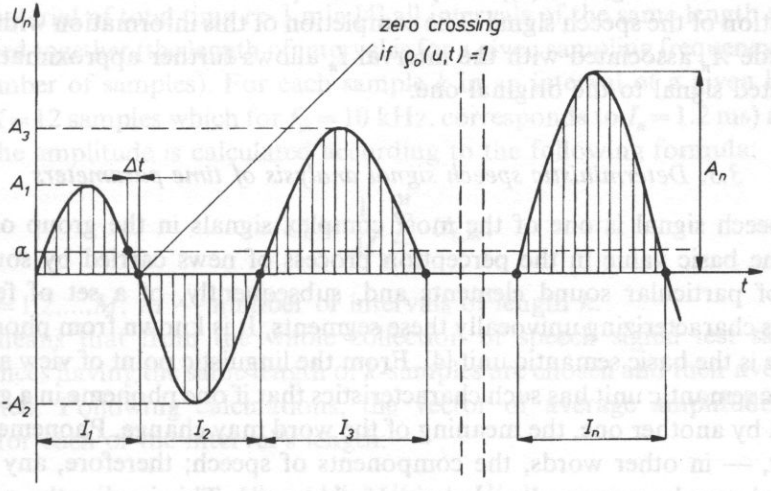

Fig. 6. The speech signal's zero crossings.

$$\rho_0(U,t)=\begin{cases}1 \text{ if } U(t) \text{ fulfilling the conditions}\\ \quad \text{of the inequalities (3.5) exists}\\ 0 \text{ otherwise}\end{cases} \tag{3.4}$$

$$\left.\begin{array}{l}U(t)\,U(t-\tau)<0\\ |U(t)|\geqslant\alpha,\ |U(t-\Delta t)|<\alpha\\ U(t)<\alpha, \text{ for } t<t<t+\Delta t\end{array}\right\} \tag{3.5}$$

where:

α — describes a certain threshold value ($\alpha < 0$) protecting from occurrence of any additional zero-crossing caused by distortions, Δt — the displacement of a zero-crossing position resulting from the fact that $\alpha \neq 0$.

Zero-crossing analysis makes it possible to isolate a few parameters characterizing the speech signal with defined precision. These are, among others, the current zero crossing density $\rho_0(t)$ and the average zero-crossing density within the segment T:

$$\rho_0 = \frac{I}{T}, \tag{3.6}$$

where: I — number of points within the segment T with $\rho_0(U,t) = 1$ or the time intervals distribution between subsequent zero crossings $R(t)$ [4].

$$R(t) = \sum_{j=1}^{I} \delta(t - t_j), \tag{3.7}$$

where: $\delta(t)$ — delta function $j = 1,2,...,I$ — the point on the axis of time with $\rho_0(U,t) = 1$.

The time intervals (in practice they are surveyed as a number of samples of the discrete form of the signal $U(n)$) carry the most part of information that allow the reconstruction of the speech signal. The completion of this information with the value of amplitude A_n associated with the interval I_n allows further approximation of the reconstructed signal to the original one.

3.3. Deterministic speech signal analysis of time parameters

The speech signal is one of the most complex signals in the group of acoustic signals. The basic value in the perception process of news carried by sound is the function of particular sound elements and, subsequently, of a set of features or parameters characterizing univocally these segments. It is known from phonetics that a phoneme is the basic semantic unit [4]. From the linguistic point of view a phoneme as the basic semantic unit has such characteristics that if one phoneme in a given word is replaced by another one, the meaning of the word may change. Phonemes are "the atoms" or, — in other words, the components of speech; therefore, any utterance pronounced may be expressed as a chain of phonemes. This implies the necessity of examining and analysing these individual components of Polish speech. It has to be observed that in spite of the fact that phonemes make up 37 classes in Polish speech, they show a significant differentiation within each of those classes. The reasons for that are the following:

a) the differences between the class subjects,

b) the differences inside the subjects,

c) the context differentiation (phonemes' context representations, i.e., allophones).

These differences constitute a significant problem when creating a relatively simple model of signal parametrization not only in vocoder technology but also in

broadly viewed research in the range of speech analysis and recognition [4]. The results of observation quasi-stationary causes of phonemes make it possible to assume the elementary dependences $A_n = f(I_n)$ (possibly corrected on the basis of statistical research). They also allow for both the choice of shapes of time envelopes of the amplitudes for the intervals U'_n and the introductory quantization of the values A_n, I_n (and possibly the shapes) in order to minimize the transmission rate.

3.4. Statistical analysis of the signal for time parameters

In order to settle the best coding parameters I_n and A_n as well as the synthesiser's analytical data, a statistical analysis of time intervals between subsequent zero-crossings I_n was carried out. The dependences between the amplitudes A_n and the length of interval I_n were settled expressing the probability function $P(A_n)$ for the given interval I_n. The average course of U_n was estimated (the amplitude's time envelope) for the n^{th} interval along with the standard deviation for each particular time sample. The data obtained from a statistically representative sound material sample for the Polish speech made it possible to settle the most probable amplitudes and provided information on the distributions of the shape of statistical envelopes (standard deviations for samples). The analysis of statistical dependences runs according to the following algorithm: for a sound material of total time $t > 1$ min [4] all intervals of the same length are chosen and grouped together (the length of interval is for a given sampling frequency measured by the number of samples). For each sample k in an interval of a given length (for example $K = 12$ samples which for $f_{pr} = 10$ kHz, corresponds to $I_n = 1.2$ ms) an average value of the amplitude is calculated according to the following formula:

$$U'_k = \frac{1}{M} \sum_{m=1}^{M} U_{k,m}, \tag{3.8}$$

where: $m = 1,2,...,M$; M — number of intervals of length k.

This means that from the whole collection of speech signal test samples the sub-sequences having the same length of k-samples are chosen and their average value is calculated. Following calculations, the vector of average amplitude values is obtained for each of the interval's length:

$$U'_k = \{U'_{k,1}, U'_{k,2}, \ldots, U'_{k,K}\}, \tag{3.9}$$

where:

$$U'_{k,K} = \frac{1}{M} \sum_{m=1}^{M} U_{k,m}. \tag{3.10}$$

The standard deviation indicates the statistical dispersion of the values $U_{k,K}$ resulting from the differences mentioned above:

$$\delta_{k,K} = \frac{1}{M-1} \left[\sum_{m=1}^{M} U_{k,K} - M U'^{2}_{k} \right]. \tag{3.11}$$

3.6. The speech signal's resynthesis

The speech signal's resynthesis realized in a contractual "receiver" of the time vocoder (see Fig. 5) runs simultaneously with the sequel of samples from a generator set to the analyzer's sampling frequency (f_{pr}). The line receiver with a memory buffer with compensates the asynchroneity resulting from the fact that the portions of data $\{I_n, A_n\}$ or the number indicating the vector representing the pair $\{I_n, A_n\}$ measured in the analyzer is transmitted in constant time distances while their usage in time (either the numbers or the vector) depends on the value of I_n (the interval's length). This makes it necessary to store the data (A_n, I_n) and allow for a slight delay of transmissoin, which should not exceed 100 ms and should not have any influence on the quality of transmission. The synthesiser's memory contains the information, the envelope shape calculation program which would also make up these values alternating their sign as well as the reconstructed signal's output through the channel of the analog-to-digital converter and the output system. Depending on the accepted option of resynthesis, the following data may be present in the memory:

a) the data concerning the most probable values of amplitudes (in this case it is possible to transmit only the values of I_n),

b) the shape or a few settled shapes of U_n related to the ranges of length of the intervals I_n.

4. Estimation of the parameters of the time vocoder

4.1. Presentation

The concept of asynchronic coding of time parameters and their analysis became the base for the functional design of the model of the functional vocoder based on an all-purpose computer hardware with acoustic input and output and specially designed software realizing not only the estimation of the time parameters but also the time vocoder's simulation of operation. Therefore the universal option was accepted; this allowed for the realization of the analysis research program and for the time vocoder's simulation of operation along with the elements of its testing and transmission quality estimation.

4.2. Research hardware

For research purposes computer hardware was utilized making it possible, owing to the software, to settle the data for analysis and the parameters of the time vocoder. The indispensable completion to the hardware were the following (Fig. 7):

a) an analog acoustic input/output system (a microphone, amplifiers, low pass filters, a loudspeaker) ordered at ZAIPSA ITA, the workshop of the Technical University of Wrocław,

b) analog-to-digital and digital-to-analog converters' card CONVERT of the TAD-01 type along with the software for the preliminary working registration and for the speech signal's output,

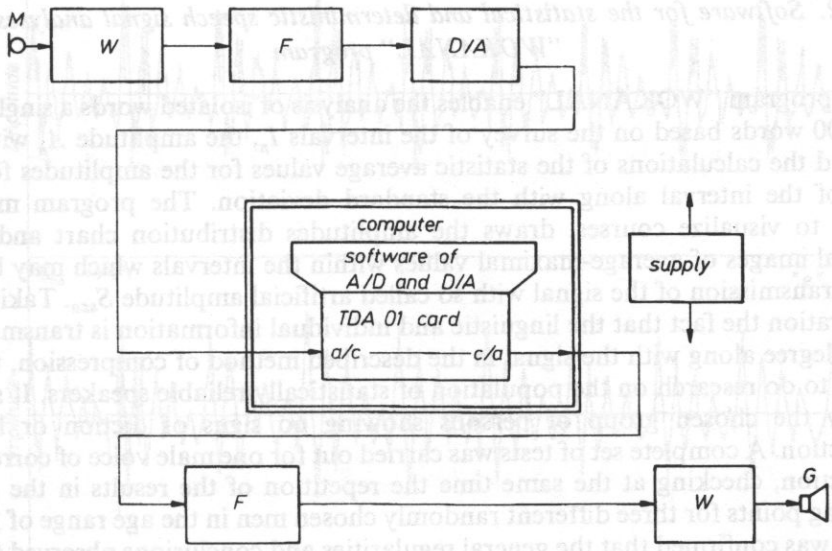

Fig. 7. The block scheme of the acoustic input and output. W — amplifier, F — filter.

c) the software for the algorithm of the analysis "WOKANAL",
d) the software realizing functions of the time vocoder "WOKCZAS".

4.3. Acoustic input/output system

The acoustic input/output system designed and developed as a part of the research program along with the A/D and D/A card is an indispensable linker between the analog and digital speech signal's structure. A block scheme of input and output systems cooperating with the A/D and D/A converter's card type TAD-01 placed on the computer's main card is shown in Fig. 7. The input system is composed of a directional microphone (M), an input amplifier, a low pass filter (F), an amplitude limiter, a registration measuring machine with an indicator. The output system is composed of a low pass filter, a tape recording amplifier, an output power amplifier with a loudspeaker. Since the sampling frequency $f_{pr} = 10$ kHz was assumed a priori, the analyzed speech signal baseband practically does not exceed 4.0 kHz. The developed systems ensure the speech signal's transmission without any distortion wsithin the frequency band limited by the low pass filter (F) from the microphone to the A/D cord and from the A/D to the loudspeaker.

4.3.1. The A/D *and* D/A *converter's card and its software*

The analog-to-digital A/D and digital-to-analog D/A converters card cooperates with both: the acoustic input and output system as well as — owing to the library functions for the Turbo Pascal compiler — the computer's compiler.

4.3.2. Software for the statistical and deterministic speech signal analyses — the "WOKANAL" program

The program "WOKANAL" enables the analysis of isolated words a single list of up to 100 words based on the survey of the intervals I_n, the amplitude A_n with their signs and the calculations of the statistic average values for the amplitudes for each length of the interval along with the standard deviation. The program makes it possible to visualize courses, draws the amplitudes distribution chart and traces statistical images of average-maximal values within the intervals which may be used for the transmission of the signal with so called artificial amplitude S_{sza}. Taking into consideration the fact that the linguistic and individual information is transmitted in a great degree along with the signal in the described method of compression, there is no need to do research on the population of statistically reliable speakers. It suffices to study the chosen group of persons showing no signs of diction or hearing imperfection. A complete set of tests was carried out for one male voice of correct and clear diction, checking at the same time the repetition of the results in the chosen examining points for three different randomly chosen men in the age range of 25–45 years. It was confirmed that the general regularities and conclusions observed for one voice may be the transmission only of the intervals I_n, i.e., the signal with the artificial amplitude. It must be considered that the statistics which allow to determine artificial amplitudes are the closest to the speaker whose utterances were analized. For other speakers this part of the observations will not be reliable.

4.3.3. Software for the process of the speech signal's analysis and resynthesis — the "WOKCZAS" program

The vocoder's software (the "WOKCZAS" program) comprises signal analysis in time, extraction of the intervals I_n and of the amplitudes A_n, coding, recording, reading and the signal's resynthesis. In the speech signal analysis part (the emitter), the intervals I_n are isolated and their length is stored. The values and signs of the amplitudes A_n for subsequent intervals are measured analogically. The program for analysis also comprises the option of extraction of the interval I_n alone (This was marked as the operation with the artificial amplitude S_{sza}). In order to ensure the same phase for the "artificial amplitude" signal and the original signal as well as to assign an amplitude with the correct sign to a given interval the appropriate sign is encoded using the same bits which are used to store the interval's length. The program encloses many options that make it possible to examine the signals in different scales, to compare the time courses, the signal's performance through the loudspeaker; the signals may also be written onto a disk or a tape recorder.

4.4. Deterministic dependences — length of time intervals' between zero-crossings and amplitude

The program "WOKANAL" makes it possible to estimate the deterministic parameters. The research was carried out on the experimental material of 37 Polish speech phonemes pronounced in the following way:

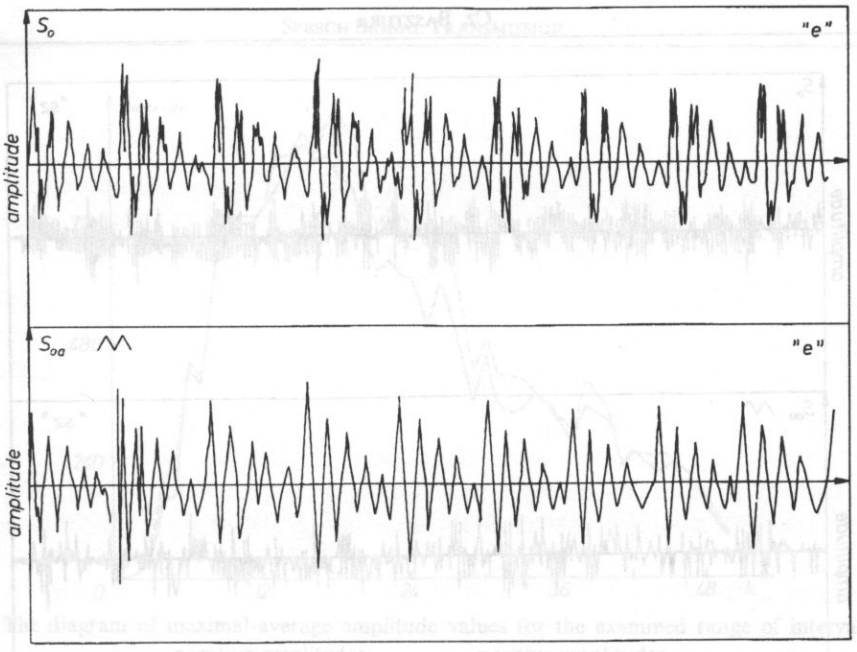

Fig. 8. The original (top) and the reconstructed (bottom) of the "e" wovel — triangular function approximation.

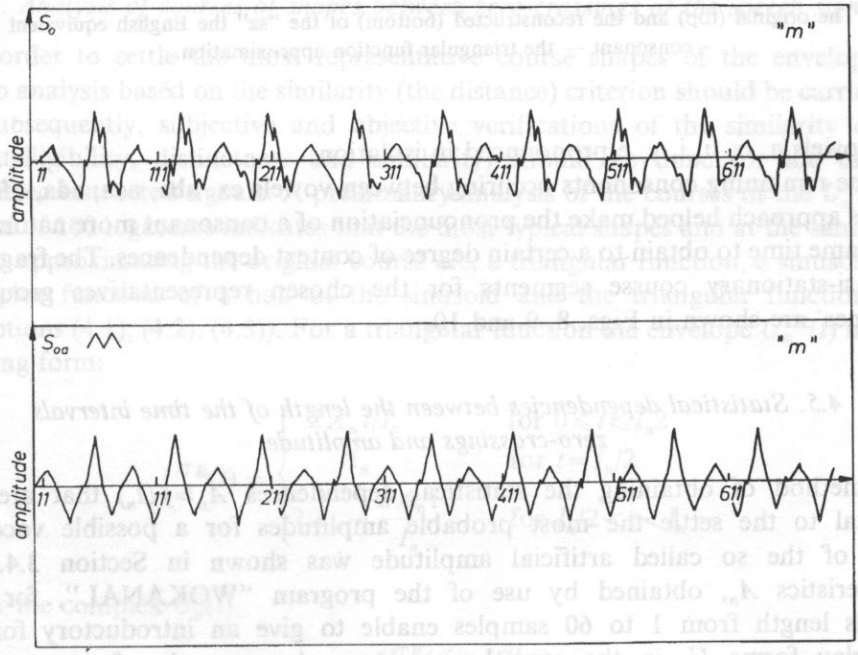

Fig. 9. The original (top) and the reconstructed (bottom) of the "m" consonant — triangular function approximation.

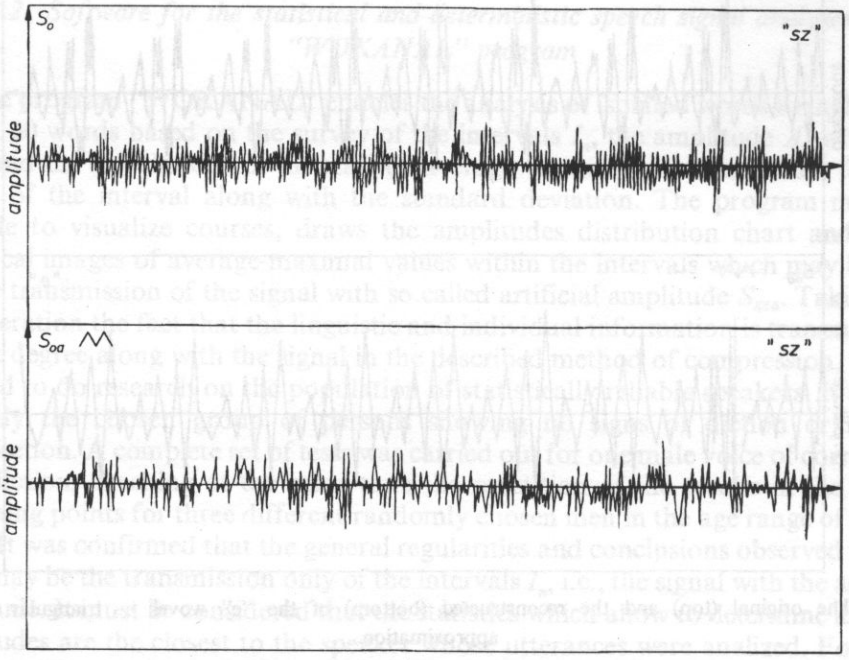

Fig. 10. The original (top) and the reconstructed (bottom) of the "sz" the English equivalent in "sh" consonant — the triangular function approximation.

- vowels/a, o, u, i, y, e/pronounced in isolation,
- the remaining consonants occuring between vowels ex. **aba, aca, ada, afa,** etc.

This approach helped make the pronounciation of a consonant more natural and at the same time to obtain to a certain degree of context dependences. The fragments of quasi-stationary course segments for the chosen representatives groups of phonemes' are shown in Figs. 8, 9 and 10.

4.5. Statistical dependencies between the length of the time intervals zero-crossings and amplitude

A method of obtaining the statistical dependences $A_n=f(I_n)$ that are fundamental to the settle the most probable amplitudes for a possible vocoder's option of the so called artificial amplitude was shown in Section 3.4. The characteristics A_n, obtained by use of the program "WOKANAL", for each intervals length from 1 to 60 samples enable to give an introductory forecast of overlay forms U_n in the resynthesis system. An example of aggregate distribution of maximum-average amplitude values for a list of 100 words is shown in Fig. 11.

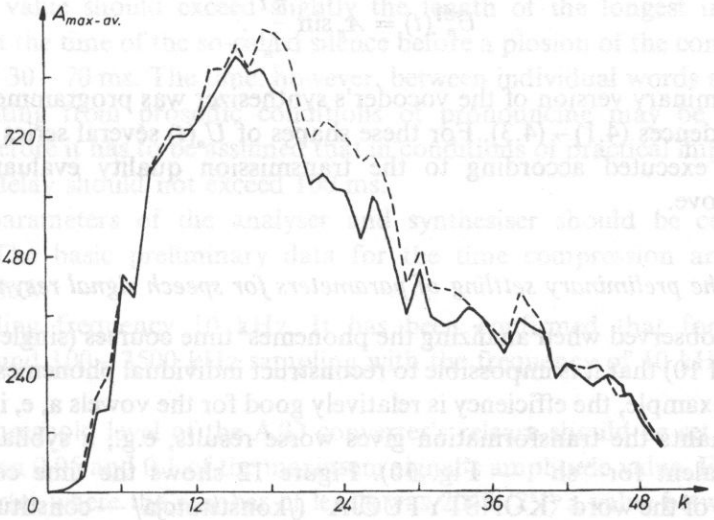

Fig. 11. The diagram of maximal-average amplitude values for the examined range of intervals. ..... negative amplitudes, _____ positive amplitudes.

4.6. Analysis of courses of shapes between zero-crossings of the speech signal

In order to settle the most representative course shapes of the envelope U_n, a group analysis based on the similarity (the distance) criterion should be carried out and, subsequently, subjective and objective verifications of the similarity degree (the intelligibility, distinctness and similarity) should be done for the original and the reconstructed signals. A preliminary analysis of the courses of the U_n of the material of 100 logatoms indicates that the most typical shapes and at the same time the best approximating the original course are: a triangular function, a sinusoid and a complex function of a half of the sinusoid and the triangular function (the assumptions (4.1), (4.2), (4.3)). For a triangular function the envelope $U_n^{TR}(t)$ has the following form:

$$U_n^{TR}(t) = \begin{cases} 2A_n t/I_n & \text{for } 0 \leqslant t \leqslant I_n 2 \\ A_n & \text{for } t = I_n/2 \\ 2A_n - \dfrac{2A_n t}{I_n} & \text{for } I_n/2 < t < I_n \end{cases} \tag{4.1}$$

and for the complex $U_n^Z(t)$

$$U_n^Z(t) = \frac{U_n^{TR}(t) + U_n^{SI}(t)}{2}, \tag{4.2}$$

where: $U_n^{TR}(kt)$ as in Eq. (4.1) dependency and

$$U_n^{SI}(t) = A_n \sin \frac{\pi t}{I_n}. \tag{4.3}$$

The preliminary version of the vocoder's synthesizer was programmed according to the dependences (4.1)–(4.3). For these shapes of $U_n(t)$ several series of measurements were executed according to the transmission quality evaluation criteria described above.

4.7. The preliminary settling of parameters for speech signal resynthesis

It has be observed when analizing the phonemes' time courses (single examples in Figs. 8, 9 and 10) that it is impossible to reconstruct individual phonemes to the same degree. For example, the efficiency is relatively good for the vowels a, e, i, y, o, u. For some consonants the transformation gives worse results, e.g., a sybilant "sz" (the Polish equivalent for "sh" — Fig. 10). Figure 12 shows the time course of the pronouncing of the word "KONSTYTUCJA" (/konstitutɕja/ — constitution). When transmitting a continuous speech or whole sentences pronounced spontaneously, delay distortions of a certain kind may arise. These distortions are caused by the necessity of a certain length of the signal buffering in the transmitter (or in the receiver) as a result of asynchronous extraction of the sequence $\{I_n, A_n\}$ connected with the variability of I_n and, therefore, with the transmitted information's matter.

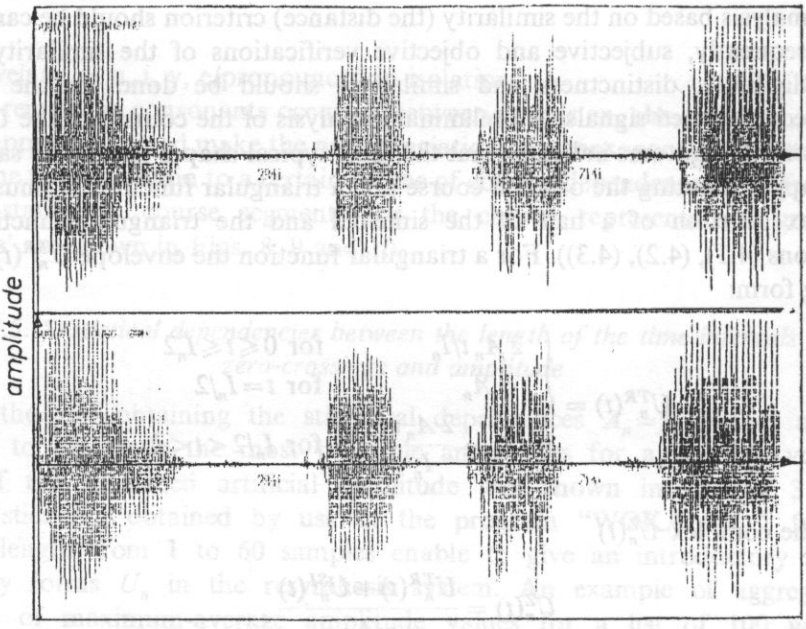

Fig. 12. The time courses for the pronounciation of "konstytucja" (constitution) /konstitutɕja/, the first window from the top — S_o, the second one — S_{oa}.

The delay's value should exceed slightly the length of the longest interval. It is assumed that the time of the so-called silence before a plosion of the consonants p, t, k, g is about 30–70 ms. The time, however, between individual words and so-called pauses resulting from prosodic conditions of pronouncing may be significantly longer. Therefore it has to be assumed that in conditions of practical implementation the time of delay should not exceed 100 ms.

Which parameters of the analyser and synthesiser should be considered as presettled? The basic preliminary data for the time compression are given and described below:

1. Sampling frequency 10 kHz. It has been confirmed that for the speech baseband 100–3500 kHz sampling with the frequency of 10 kHz is satisfactory.
2. The threshold level of the A/D converter's release should be set to the value between 0.05 and 0.1 of the maximum signal's amplitude value. For the 12-bit converter where the number of levels was 2048 (2^{11}) a value between 100 and 200 was assumed.
3. The signal's beginning retrieval was based on the verification as to whether a certain number of m samples exceed the threshold level α. It was assumed that if $m=10$ samples exceed the threshold level, this point is the beginning of the signal's analysis. The value of m and the mode of indication of the beginning as well as the end is an arbitrary matter from the point of view of the method and was not subject to any analysis.
4. Three shapes constituted of shape analysis of the envelope of the signal $U(n)$ assigned to the intervals in the synthesiser irrespective of the value of the amplitude S_{oa}:
   - a triangle,
   - a sinusoid,
   - average superposition of a sinusoid and a triangle.

   The dependences for the shapes mentioned above were described by the patterns (4.1), (4.2) and (4.3). The results of formal and informal audio monitoring research showed that "statistically" the best quality of sound for the listeners as well as the best distinctness and intelligibility are ensured by the complex function of a sinusoid and a triangle. Examples of the approximation of the vowels "a" and "o" by the three shapes: a sinusoid, a triangle and a complex function are shown in Figs. 13, 14 and 15.
5. The variability range of values of the intervals I_n with amplitudes $A_n>0$ was limited from 0.1 to 6 ms. The greater lengths are represented as silence, plosion distances, etc. This corresponds to the number of samples from 1 to 60 for $f_{pr}=10$ kHz.

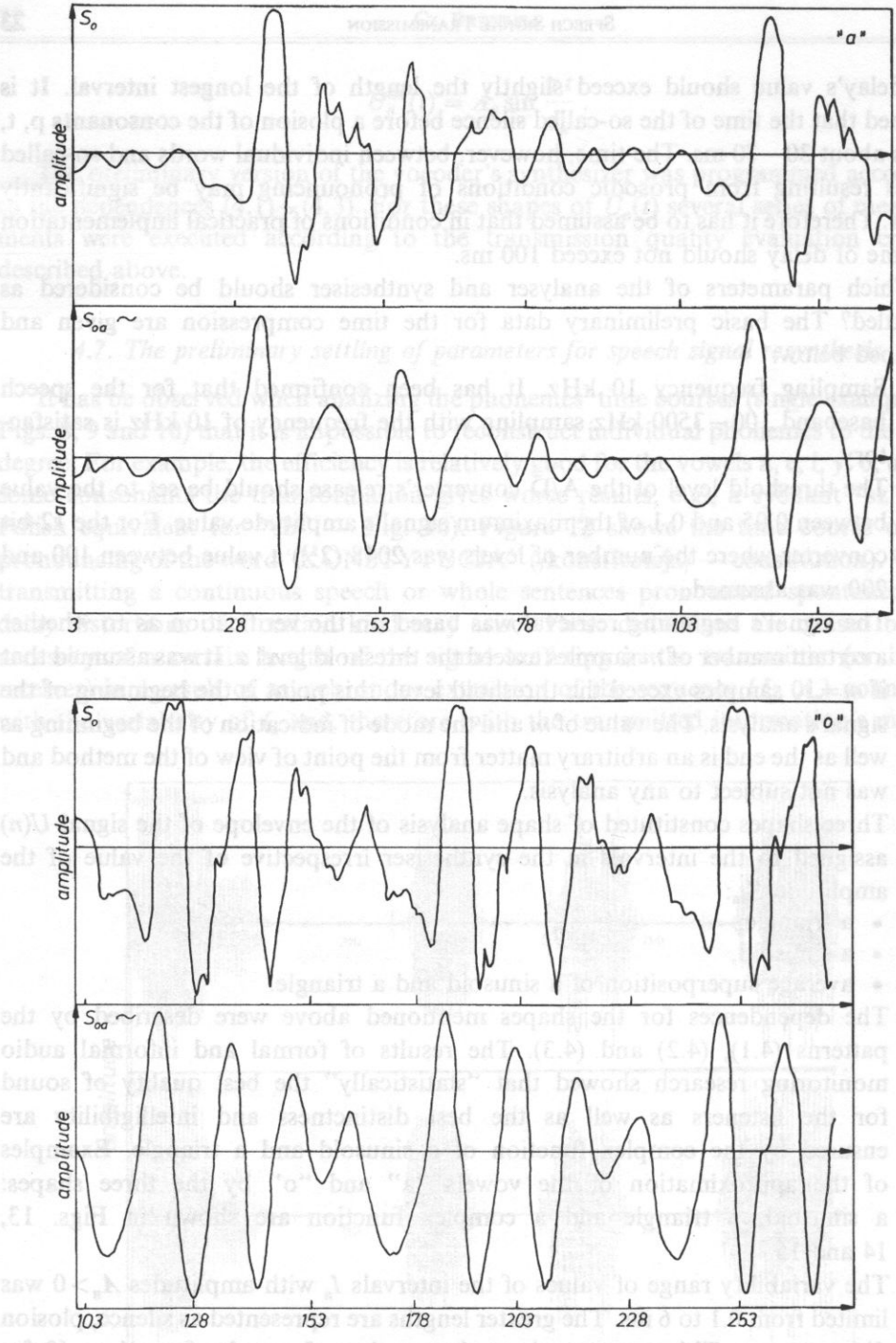

Fig. 13. The time segments of the "a" and "o" vowels approximated by a sinusoid function (the assumption (4.3)).

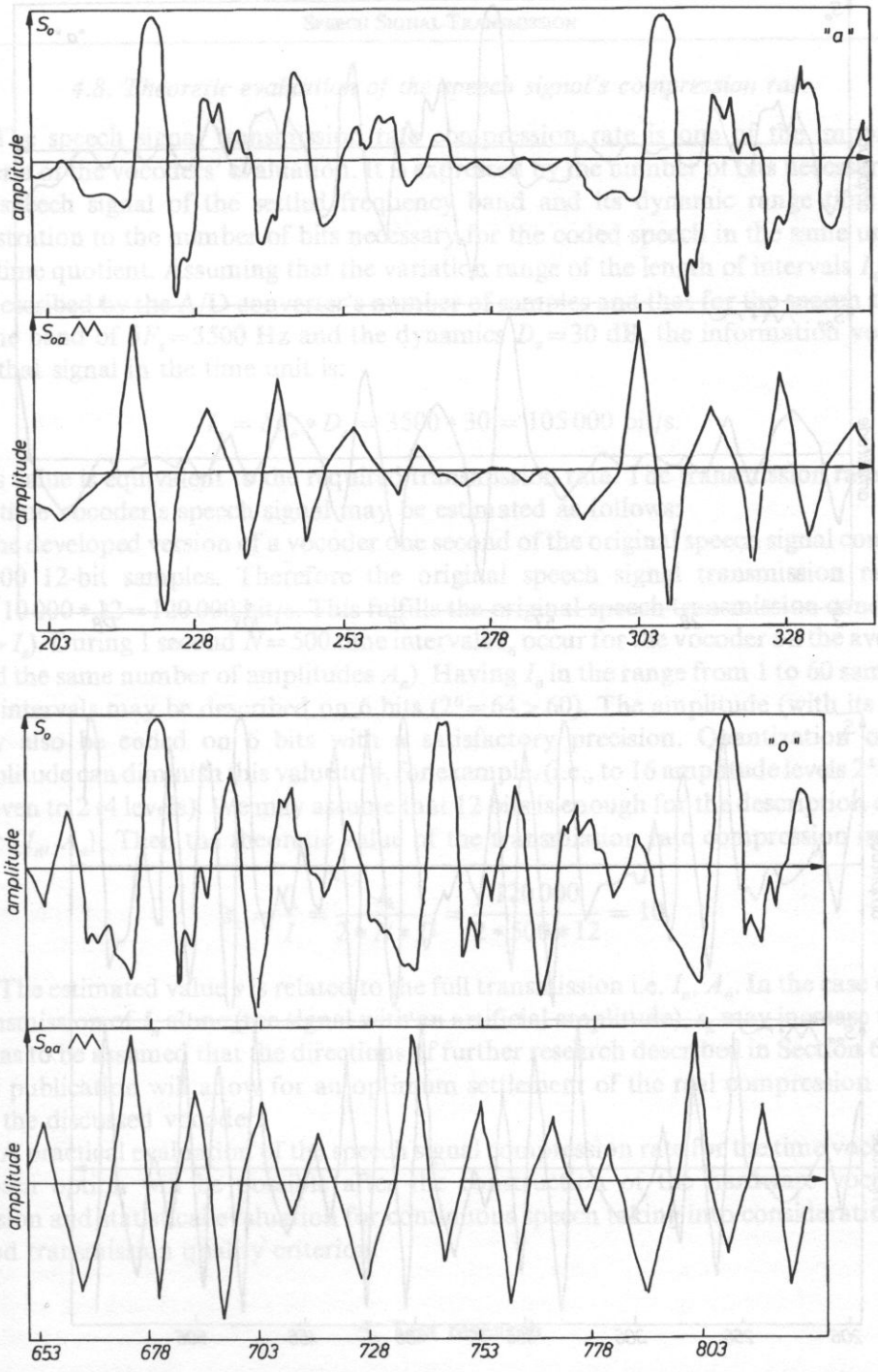

Fig. 14. The time segments of the "a" and "o" vowels approximated by a triangular function (the assumption (4.1)).

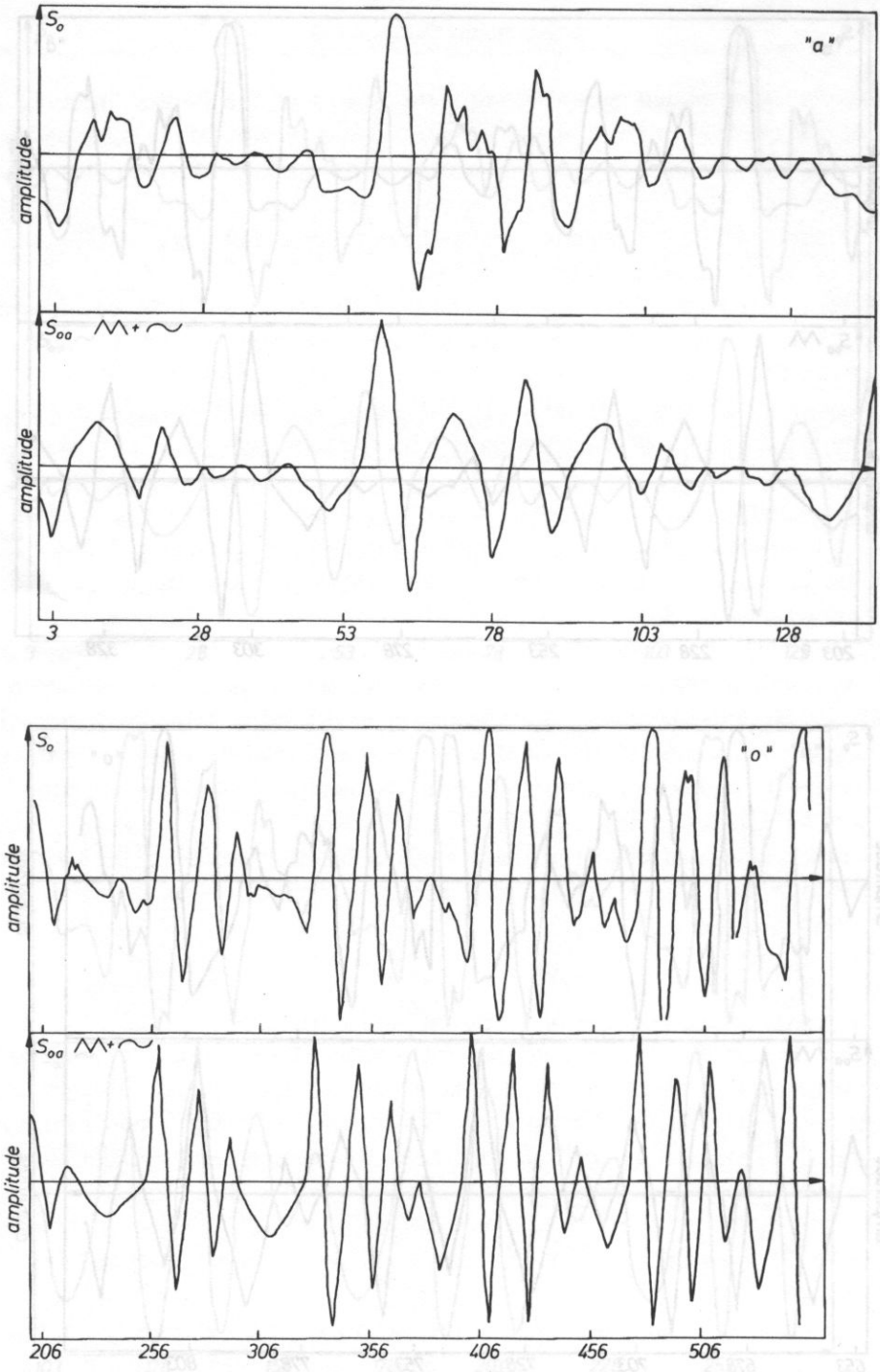

Fig. 15. The time segments of the "a" and "o" vowels approximated by a complex function (the assumption (4.2)).

4.8. Theoretic evaluation of the speech signal's compression rate

The speech signal transmission rate compression rate is one of the important criteria of the vocoders' evaluation. It is expressed by the number of bits necessary for the speech signal of the settled frequency band and its dynamic range time unit registration to the number of bits necessary for the coded speech in the same unit of the time quotient. Assuming that the variation range of the length of intervals I_n may be described by the A/D converter's number of samples and that for the speech signal in the band of $\delta F_s = 3500$ Hz and the dynamics $D_s = 30$ dB, the information volume for that signal in the time unit is:

$$I_s = \delta F_s * D_s = 3500 * 30 = 105\,000 \text{ bit/s.}$$

This value is equivalent to the required transmission rate. The transmission rate I_n of the time vocoder's speech signal may be estimated as follows:
In the developed version of a vocoder one second of the original speech signal contains 10 000 12-bit samples. Therefore the original speech signal transmission rate is $I_k = 10\,000 * 12 = 120\,000$ bit/s. This fulfills the original speech transmission condition ($I_k > I_s$). During 1 second $N = 500$ time intervals I_n occur for the vocoder on the average (and the same number of amplitudes A_n). Having I_n in the range from 1 to 60 samples, the intervals may be described on 6 bits ($2^6 = 64 > 60$). The amplitude (with its sign) may also be coded on 6 bits with a satisfactory precision. Quantization of the amplitude can diminish this value to 4, for example, (i.e., to 16 amplitude levels $2^4 = 16$) or even to 2 (4 levels). We may assume that 12 bits is enough for the description of the pair $\{I_n, A_n\}$. Then the theoretic value of the transmission rate compression is:

$$s_k = \frac{I_k}{I} = \frac{I_k}{2 * N * D} = \frac{120\,000}{2 * 500 * 12} = 10.$$

The estimated value s is related to the full transmission i.e. I_n, A_n. In the case of the transmission of I_n alone (the signal with an artificial amplitude), s_k may increase to 20. It has to be assumed that the directions of further research described in Section 6.2. of this publication will allow for an optimum settlement of the real compression value for the discussed vocoder.

A practical evaluation of the speech signal compression rate for the time vocoder's chosen option will be possible after the construction of the hardware vocoder's version and statistical evaluation for continuous speech taking into consideration the good transmission quality criterion.

5. Test research

5.1. Purpose and methodology of the research

The diversity of transmission quality estimation methods and techniques used for digital systems for coding and the transmission of information is a great obstacle

when comparing directly the quality of such systems. It has been decided to consider the following methods as the estimation criteria for the time vocoder operating according to the newly-introduced method of transmission rate compression:

A. Subjective measurements:
- average logatom distinctness — SWL,
- average word intelligibility — SZW,

B. Objective measurements:
- mean-square difference,
- cross corelation,
- sonographic analysis.

The essential purpose of the research mentioned above was the following:
- the verification of correctness of the preliminary choice of the time parameters' $\{A_n, I_n\}$ range of variability. (see Section 4.7 of this publication).
- the statistical objectivized transmission quality estimation expressed in the subjective quality scale (SSJ), logatom distinctness (SWL) or word intelligibility (SZW) as well as in objective measures of the original signal's $S_0 - \{U_o(t)\}$ conformity with the reconstructed signal $S_{oa} - \{U_{oa}(t)\}$.

5.2. Measurements of the time vocoder's distinctness and intelligibility

The correct sound material assortment for subjective (acoustic) estimation tests is a problem of particular significance. The assumption that the test material represents a sound production which is most characteristic of the predicted applications of the discussed teletransmission system should be the basic rule of the test material's assortment [23].

The method of logatom and word lists phonetically and structurally compensated is one of the recommended test types. Phonemes and succession structures are determined in agreement with the proportions in natural speech [23]. Another question in the correct process of subject research is the choice and training of the audio monitoring staff. According to ISO recommendations [25], a person's hearing is correct if its hearing threshold value does not exceed 10 dB for any test frequency less than 4000 Hz and 15 dB within the band from 4000 Hz to 6000 Hz. It is recommended that the audio monitoring group be large enough so that the average results do not change much along with the increase of the group's population. It has been stated that such a minimal number of the group's population is 5 persons of each sex. The mode of the measurements' process preceded by an adequate training was consistent with ISO recommendations [23]. The calculation and the statistical inference modes were also executed according to the rules of statistical methods. The signal with the original amplitude was (S_{oa}) subject to the subjective research. The results (in %) of obtained distinctness and intelligibility along with standard deviations for the approximation using the complex function are shown in Tables 2 and 3.

The results of intelligibility and distinctness tests indicate a good quality of transmission of the speech signal with the original amplitude. This means that the

Table 2. Preliminary distinctness test (SWL) results

The listener	S_{oa} [%]
1	68
2	62
3	70
4	63
5	62
W_L [%]	65
δ_{SWL} [%]	7.48

Table 3. Preliminary results of the intellibility test SZW and the listener's subjective evaluation SOS scaled from 1 to 5

The listener	S_{oa}	
	[%]	SOS (1–5)
1	88	4
2	90	4
3	89	4
4	93	4.5
5	87	4
SZW [%]	89.4	4.2
δ_{SZW} [%]	4.2	0.89

amplitude information transmission is necessary for a good transmission quality. It is known from the speech signal's analysis [4] that this may be slow-changing information which allows further compression and significant improvement of the transmission's quality, close to S_{oa}. The obtained results, the best ones for the complex function (a triangle and a sinusoid) approximating the amplitude, make it possible to obtain the transmission's quality estimated by the listeners (SOS) and given by the results of the SLW and SWZ measurements as good. The preliminary subjective measurements were carried out on a representative sound material which, however, was limited by time consumption. Therefore the obtained results should be considered as preliminary ones indicating the range for the future implementation of the method.

5.3. *Objective measurements*

The tests of objectivized comparisons of the original signal S_o with the reconstructed signals S_{oa} with the original amplitude were carried out utilizing the following:

a) the mean-square distance,

b) the standardized cross-corelation the cross corelation coefficient

c) the sonographic spectrum evaluation.

a) The mean-square difference. The results of R_{sk} for comparisons of the determined speech signal's segments are shown in the Table 4. The mean-square difference is a measure of similarity between the sampled signal courses the original S_o

Table 4. The mean-square difference results for the speech signals' segments (phonemes) of 25 ms length

No.	phonemes	R_{sk} (S_o, S_{oa})
1	/ a /	12277
2	/ e /	32204
3	/ m /	31267
4	/ z /	14232
5	/ ʃ /	62944

and the reconstructed S_{oa}. Owing to the different characteristics of the speech sound courses, the comparative analysis should be limited to the same phonemes. In Table 4 different values of R_{sk} for S_{oa} show different "compatibility" of the signals with the original amplitude that is sent along with the original signals S_o. Not should be made attention of the dispersion of proportions among the values R_{sk} S_o, S_{oa}. This results directly from time structure differences of the phonemes.

b) The cross-correlation. In Table 5 the values of the cross-correlation factor r_{uv} are shown for the chosen segments of the quasi-stationary signals $t=25$ ms. As in the case of the mean-square

Table 5. The cross-correlation factor values

No.	phonemes	$r_{uv} - (S_o - S_{oa})$
1	/ a /	0.91
2	/ e /	0.85
3	/ m /	0.57
4	/ z /	0.88
5	/ ʃ /	0.18
average		0.88

difference, a significantly big dispersion of the cross-correlation factor values was observed for different letters. For the most part the best cross-correlation have the wovels and the sonant consonants. The worst was obtained for sibilants.

c) Sonographic analysis. An example of a speech signal 3-dimensional representation is shown in Fig. 16 (/riba/ — fish) as a spectrogram of the word "RYBA". The

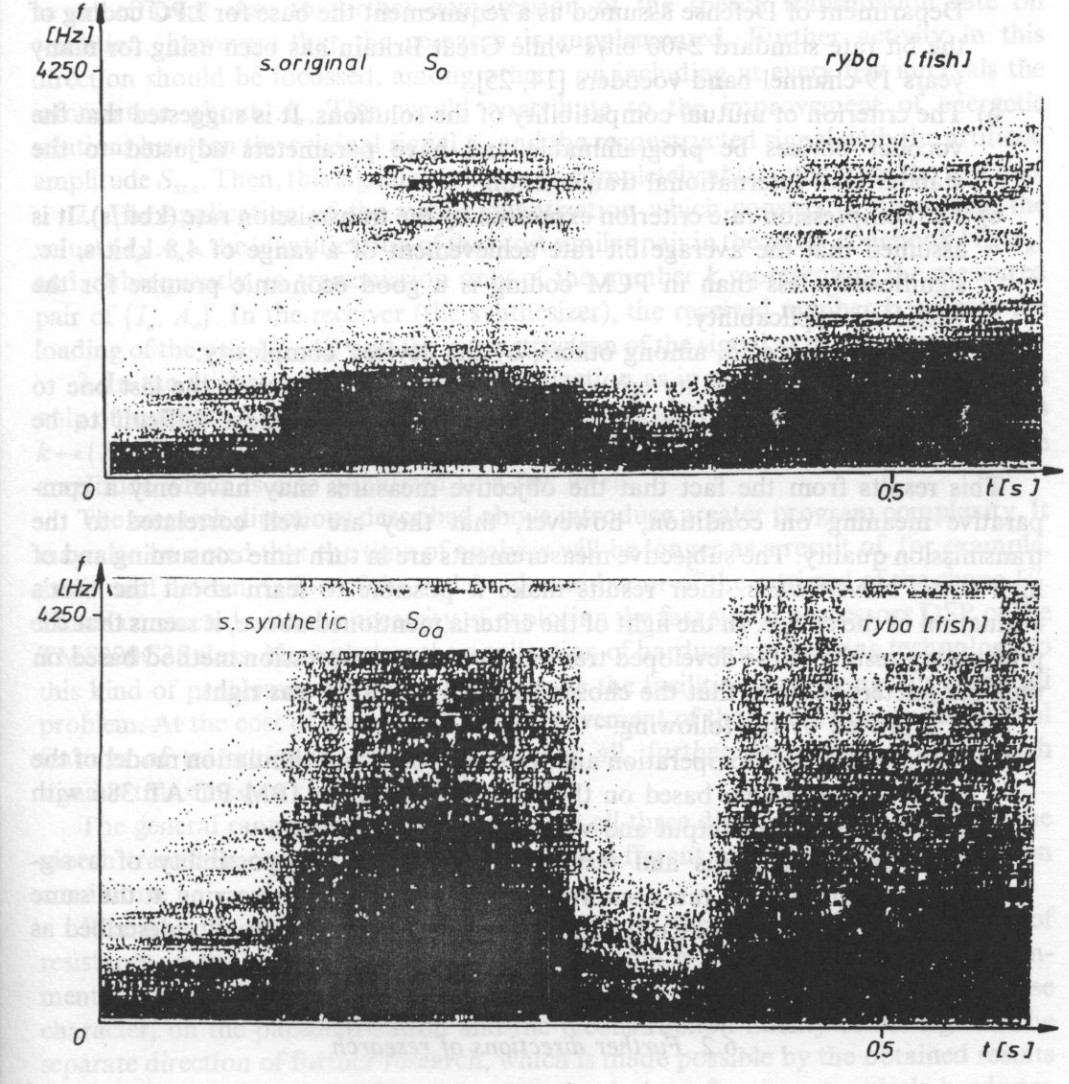

Fig. 16. The spectrogrammes of the pronounciation of "Ryba" (/riba/ — fish) from the top S_o, S_{oa}.

sonographic pictures of the signals confirm the pattern similarity of the reconstructed speech S_{oa} (the bottom spectrogram) with S_o (the upper spectrogram).

6. Summary and conclusions

6.1. Summary

The application of specific parametric coding solutions in vocoders is conditioned by some criteria. The following may be considered as the main ones:

a) The criterion resulting from the specific demands. For example, the US Department of Defense assumed as a requirement the base for LPC coding of the bit rate standard 2400 bit/s while Great Britain has been using for many years 19-channel band vocoders [14, 25].

b) The criterion of mutual compatibility of the solutions. It is suggested that the vocoder devices be programmable and have parameters adjusted to the standards of international transmission.

c) The compression rate criterion expressed by the transmission rate (kbit/s). It is assumed that the average bit rate achievement of a range of 4.8 kbit/s, i.e. significantly less than in PCM coding is a good economic premise for the vocoder's applicability.

d) The costs expressed, among others, by the devices' complexity.

e) The transmission quality criterion. Although this criterion is the last one to be mentioned, it is one of the most important and most difficult to be verified.

This results from the fact that the objective measures may have only a comparative meaning on condition, however, that they are well correlated to the transmission quality. The subjective measurements are in turn time consuming and of high costs. Nevertheless, their results make it possible to learn about the user's evaluation of the device. In the light of the criteria mentioned above, it seems that the preliminary results of the developed transmission rate compression method based on time parameters confirm that the chosen research direction was right.

This is proved by the following:

- the developed, set to operation and tested introductory simulation model of the time vocoder system based on the universal computer IBM PC AT 386 with acoustic input and output and with the proper software,
- the vocoder's analysis and test results indicating the possibility of a significant transmission rate compression, up to 10 times, preserving at the same time the transmission quality falling into the range of the quality described as 4 (good).

6.2. Further directions of research

The developed transmission rate compression method based on the extraction of the time parameters I_n and A_n does not fall back into its basic structure described in this publication, what is an obvious advantage of the method. The performed research and analyses proved that modifications of this method enabling, first of all, further speech signal transmission rate minimalization are possible. Three approaches

may be distinguished, which were mentioned indirectly in the previous sections of this publication:

1. A transmission only of the lengths of the intervals I_n (a signal with an artificial amplitude S_{sza}).

The preliminary analysis of the transmission only of I_n, which was taken into consideration in the "WOKCZAS" program, makes it possible to claim that this may be an efficient way to further compression of the speech transmission rate on condition, however, that the research is supplemented. Further activity in this direction should be focussed, among others, on including at every few intervals the information about A_n. This would contribute to the improvement of energetic relations between the original signal S_o and the reconstructed signal with the artificial amplitude S_{sza}. Then, this signal would not be completely abstracted from the real one.

2. The application of the vector quantization which consists in measuring the values $\{I_n, A_n\}$, the identification of the most similar pair in the pairs' codebook $\{\hat{I}_n, \hat{A}_n\}$ and, subsequently, in transmission only of the number k representing the measured pair of $\{I_n, A_n\}$. In the receiver (the synthesizer), the received number k initiates the loading of the pair $\{I'_n, A'_n\}$ and the reconstruction of the signal as in the basic method.

3. In the utilization of the vector quantization as in point 2: — the transmission only of numbers k of the intervals I_n codebook and a few estimated shapes K_n^i that is $k \leftrightarrow \{I'_n, K_n^i\}$, as well as the information transmission at every m interval about the amplitude which is also quantized to a few levels.

The research directions described above introduce greater program complexity. It may also be noted that the time of analysis will be longer as a result of, for example, additional determination of the codebook number or of the retrieval of the shape U'_n. This is connected with the necessity of exploring the fast signal processors DSP of the TMS320C25 class. Nevertheless the application of hardware processing technology to this kind of problems nowadays may provide the facilities for solving this technical problem. At the cost of complexity the improvement of the sound quality, individual features of transmission fidelity and, above all, further decreasing of the speech signal's transmission rate may be expected.

The general range of research, assigned to all three directions, should enclose the speech transmission quality evaluation under different auditorial and transmission conditions.

Research, in particular, should be focussed on the examination of the rate of resistance to the signal's distortions in the transmitter's and the receiver's environments and on the methods of diminishing the distortions' influence, mainly of a noise character, on the parametrization and the reconstruction fidelity of the signal. The separate direction of further research, which is made possible by the obtained results and the developed "tools" for research is the design of a time vocoder's hardware implementation.

This project was developed on the basis of The Research Project realized on request of the Ministry of National Education, CPBR, GRT-2.

References

[1] D.R. ALLEN, W.J. STRONG and E.P. PALMER, *Experiments on the intelligibility of low frequency speech codes*, JASA **70**, 5, 1248–1255 (1981).

[2] А.Д. АРХОПОВА, М.А. САПОЖКОВ, *О качестве вокодерной речи*, Акустический Журнал, **16**, 3, 345–353 (1970).

[3] T.P. BARNWELL, *Objective measures for speech quality testing*, J. Acoust. Soc. A., **66**, 6, 1658–1663 (1979).

[4] Cz. BASZTURA, *Sources, signals and acoustical images — processing, analysis and recognition* [in Polish], WKiŁ, Warszawa 1988.

[5] R.E. CROCHIERE et al., *A study of objective measures for speech waveform coders*, Murray Hill, New Jersey.

[6] W.R. DAUMER, *Subjective evaluation of several efficient speech coders*, IEEE Trans. Communic. **COM-30**, No 4 April 1982, 665–662 (1982).

[7] W.R. DAUMER and J.R. CAVANAUGH, *A subjective comparison of selected digital coders for speech*, Bell Syst. Techn. J. **57**, 9, 3109–3165 (1978).

[8] J.A. FELDMAN, E.M. HOFSTETTER, M.L. MLPASS, *A compact, flexible LPC vocoder based on a commercial signal processing microcomputer*, IEEE Trans. on Acoustics, Speech and Signal Processing, vol. **ASSP-31**, 1, 252–257 (1983).

[9] B. FETTE, D. HARRISON, D. OLSON and S.P. ALLEN, *A family of special purpose microprogrammable digital signal processor IC's in an LPC vocoder system*, IEEE Trans. on Acoustics, Speech and Signal Processing, vol **ASSP-31**, 1, 273–281 (1983).

[10] R. GUBRYNOWICZ, *Application of zero-crossing method in speech signal analysis and automatic word recognition*, (in Polish) IFTR PAS Report 27 (1974).

[11] D.J. GOODMAN and R.D. NASH, *Subjective quality of the same speech transmission conditions in seven different countries*, IEEE Trans. Commun., **COM-30**, 4, 642–654 (1982).

[12] D.J. GOODMAN, J.S. GOODMAN and M. CHEN, *Intelligibility and ratings of digitally coded spedch*, IEEE, vol. **ASSP-28**, 5, 403–408 (1978).

[13] A.H. GRAY and D. MARKEL, *Distance measure for speech processing*, IEEE Trans. Acoust. Speech and Sign. Proc., **ASSP-24**, 380–391 (1976).

[14] N. KITAWSKI, M. HONDA and K. ITOH, *Speech-quality assessment method for speech-coding system*, IEEE Communications Magazine, **10**, 26–33 (1984).

[15] B.F. LOGAN, *Signals designed for recovery after clipping — localization of infinite products*, ATT Bell Laboratories Technical Journal, **63**, 2, 261–285 (1984).

[16] P. MERMELSTEIN, *Evaluation of segmental SNR measure as an indicator of the quality of ADPCM coded speech*, JASA **66**, 6, 1664–1667 (1979).

[17] В.Ф. МИХАЙЛОВ, Л.В. ЗЛАТОУСТОВА, *Измерение параметров речи*, Радио и Связ, Москва 1987.

[18] M. NAKATSUI, *Subjective SNR measure for quality assessment of speech coders a cross language study*, MAT., ICA **12 A1–1**, Toronto 1986.

[19] М.В. НАЗАРОВ, Ю.Н. ПРОХОРОВ, *Методы цифровой обработки и передачи речевых сигналов*, Радио и связ, Москва 1985.

[20] A.V. OPPENHEIM [Ed], *Application of digital signal processing*, MIT, Cambridge, Mass., 02139, Prentice-Hall Inc., USA 1978.

[21] N. OSAKA and K. KAKEHI, *Objective model for evaluating telephone transmission performance*, Review of the Electrical Communications Laboratories, **34**, 4 (1986).

[22] Н.Т. ПЕТРОВИЧ, М.В. КАБЛАКОВА, Н.И. КОЗЛЕНКО, *Передача сигналов методами КИМ–ОФТ*, Радио и связ, Москва 1985.

[23] *Draft Proposal* ISO/DP 4870, *Recommended methods for intelligibility tests.*

[24] K. REEDER, W.J. STRONG and E.P. PALMER, *Preliminary study of a low frequency formant based speech code for the several hearing impaired*, JASA, **61**, 5, 1379 (1977).

[25] М. САПОЖКОВ, М. МИХАЙЛОВ, *Вокодерная связ*, Радио и связ, Москва 1983.

[26] M.E. SMITH, K.E. ROBINSON and W.J. STRONG, *Intelligibility and quality of linear predictor and eigen parameter coded speech*, IEEE, vol. **ASSP-29**, 3, 391–395 (1981).
[27] W. SOBCZAK [Ed.], *Problems of teleinformatics*, WKiŁ, Warszawa, 1984.
[28] S. WANTANABE and K. INOMOTO, *On a loudness function of artificial speech*, Acustica, **44**, 4 (1980).
[29] W. WOLFF, *Untersuchungen zur subjectiven Qualitat vom Spektrum des uberlagerten Rauchnes*, Frequenz, **34**, 6–9 (1980).
[30] А.Г. Зюко и другие, *Помехоустойчивость и ефективность систем передачи информации*, Радио и связ, Москва 1985.