voice recognition, word detection,
isolated words

Piotr PORWIK*, Robert PROKSA*

WORD EXTRACTION METHOD IN HUMAN SPEECH PROCESSING

A major problem in isolated-word speech recognition systems is detection of the beginning and ending boundaries of the word. It is an essential of speech recognition algorithms, where signal speech segments should be reliably separated. During speech recognition background noise is also recorded, hence the word isolation is difficult. The parametric representation of the speech must provide enough information to characterize the words and to differentiate between acoustically similar words. In this paper the method of words extraction from human speech will be considered.

1. INTRODUCTION

Speech recognition is one of the desired supporting technologies used after voice recognition, as a natural method of the voice commands generation. For this reason computer control by means of human voice is very attractive for many users. Speech–recognition programs do not understand what words mean, but isolated words can be recognised and used as a context tool. This information helps the computer with choosing the most likely word from the database [8,7]. An important problem in the recognition system of isolated words is the determination of upper/lower limits of the uttered word, normalization and the time of word duration. These procedures allow to reduce a number of bytes of stored data. In addition, computation time can be also significantly reduced. If an uttered word can be isolated (is appropriately recorded in the so-called time window), then preliminary word detection can be omitted. In cases when the speech signal is continuous, some parameters have to be determined and the boundaries of the uttered word should be detected. Detection of the word boundaries can be carried out on the basis of the signal energy determination [5].

Determination of the word boundaries is difficult because during speech many sources of noise can be registered, for example surroundings noise, noises during breathing-in, breathing-out, influence of low energy sound in some speech phases, sudden speech breaks and so-called reflexive sound articulation [1]. The location of the beginning and end of an isolated utterance in an environment without noise may be determined by the use of a simple energy threshold. A clean environment is one in which the energy of the speech sounds, such as weak fricatives, is significantly greater than the background energy. In realistic problems there are troubles in finding the best estimation of the beginning and end points of the spoken word in the presence of any background noise.

In this paper the two thresholds have been used. The energy and power threshold allow estimating boundaries of the single word in the human speech process.

2. VOICE ACQUISITION

Continuous voice acquisition and automatic voice recognition system can be applied if during voice recognition suitable sound level (threshold level) is guaranteed. All acoustic signals over threshold level can be interpreted as uttered Works. If acoustic signals are below threshold level, then sounds are interpreted as silence or noise. In this paper the appropriate sound threshold level is characterized by means of the parameter *Pd*. Elimination of the background noises can be realized automatically. In the first stage, the system is activated and during 1*s* silence signal is recorded and stored in the computer memory. The appropriate threshold level is determined from formula:

$$Pd = \frac{\sum_{i=0}^{L} s_i}{L} + b \qquad (1)$$

where:

L – number of the sound samples,

s_i – value of the t-th sample, i=1,…,L,

b – constant (in our case $b = 4$).

or with the aid of the formula:

$$Pd = \max(s_i) + b \qquad (2)$$

* Institute of Informatics, Silesian University, Będzińska 39, 41-200 Sosnowiec, Poland e-mail: piotr.porwik@us.edu.pl; proksa@o2.pl

A constant *b* eliminates registered by microphone additional noises. For small values of the constant *b*, preliminary word detection is more difficult because the microphone should be very sensitive. Between words there are silence intervals. Time duration of the last interval, after uttered word, should be precisely fixed.

In cases if length (in seconds) of the last interval is long, then such a word is recognized as the uttered word and the next word is awaited. In some cases appropriate determination of the length of intervals between uttered words can be difficult. Such cases will appear for low energy sounds (Fig.1). In Polish language it occurs for sound "s" for example, where attack and delay-sustain-release phases can be wrongly recognized as silence intervals [8]. Similar troubles will occur for words with low-energy phonemes (so-called weak fricatives). In such a case, a correctly uttered word can be split by recognition procedures into two or more parts.

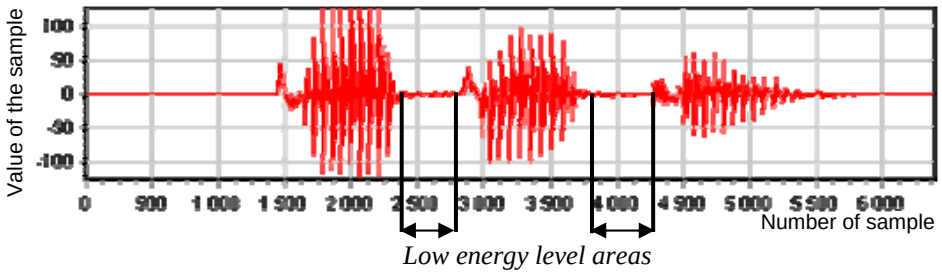

Fig.1 Word „property", incorrectly split into 3 parts

In our approach, at the beginning of the recorded sound there are also recorded samples, which occurred directly before the uttered word. In such case, low energy sounds, occurred at the speech beginning are preventively saved. Prepared data can be analyzed in recognition system of isolated words.

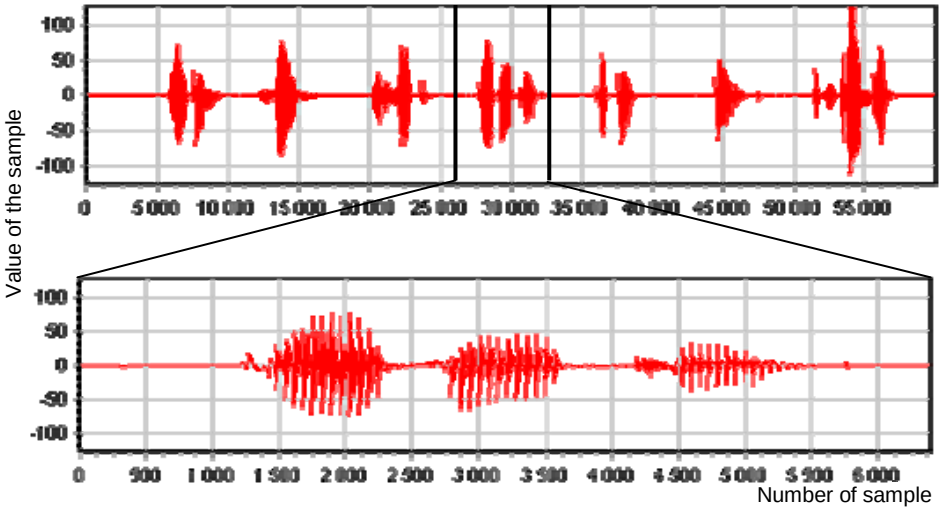

Fig.2 Preliminary detection of the word "property"

3. FILTERING

Recorded speech signals are filtered by means of the pre-emphasis filters. This operation gives then new scale of the signal – are amplified components of the high frequencies of the speech signal (Fig.3) [9,2,6]. In many cases the FIR-type filters are used (*Finie Impulse Response Filter*). It is the first order high-pass filter and can be simple realized by means of integrated circuit or by software approach [3,4,6]. Operation of the simple FIR filter can be described by equation:

$$s_i' = s_i - a s_{i-1} \tag{3}$$

where:

s_i' – digital signal of the *i*-th sample after FIR filtering,

s_i – digital signal of the *i*-th sample before FIR filtering,

a – filtering coefficient. For speech signals a=0.937.

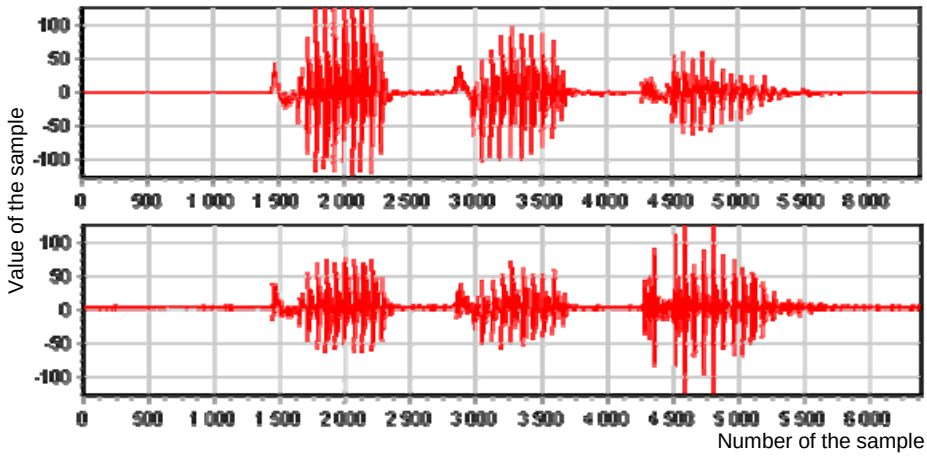

Fig.3 The word "property" after and before FIR filtering

It should be noticed that FIR procedures can be realized by means of the specialised library functions in many programming languages.

4. FRAME SIGNAL

In the proposed solution speech signal is divided into frames. The frame can be characterized by means of equation:

$$s_t(n) \equiv s(n - t \cdot M), \quad 0 \le n < D, \quad 1 \le t \le T \tag{4}$$

where:

$s_t(n)$ – value of the n-th sample of signal, for frame number t,
D – length of the frame (characterized as number of samples),
T – number of the frames,
M – displacement – used if frames are overlapped.

Signal partition has large influence on the word detection quality. If frames are too short – computation time is long. If frames are too long – word detection quality can be very low. Let:

$$L_s \bmod L_F = R \tag{5}$$

where:

L_S – length of the recorded sound (in samples),
L_F – length of the frame (in samples).

then:

$$R = \begin{cases} 0 & \text{word was appropriately recorded} \\ C & \text{word was inappropriately recorded} \end{cases}$$

In case when R=C, recorded signal is cut of C/2 samples at the beginning and at the end (Fig. 4). For this reason all frames have the same length.

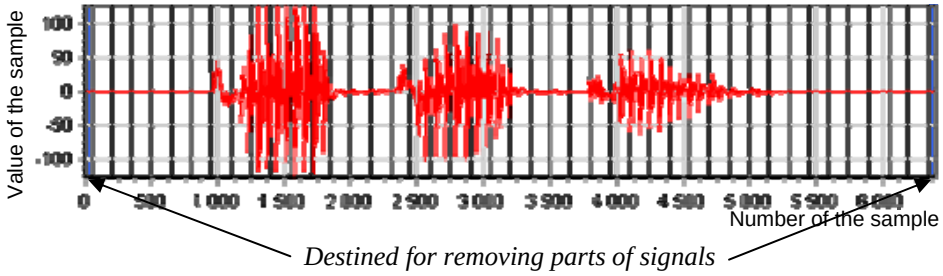

Fig.4 Signal frames which should be removed

5. FRAME ENERGY

In the proposed solution speech signal is divided into frames. For each frame, energy of the signal is computed. It allows us to detect places of silence and sounds in signal.

$$E(l) = \sum_{n=1}^{N} s_l^2(n) \ , \qquad 1 \le l \le K \tag{6}$$

$$P(l) = \log_{10}(E(l)) \qquad 1 \le l \le K \tag{7}$$

where:

l – number of current frame,
K – number of all frames,
N – number of samples in the frame,
$s_l(n)$ – value of the n-th sample in the frame l,
$P(l)$ – power of signal in the frame l,
$E(l)$ – energy of the l-th frame.

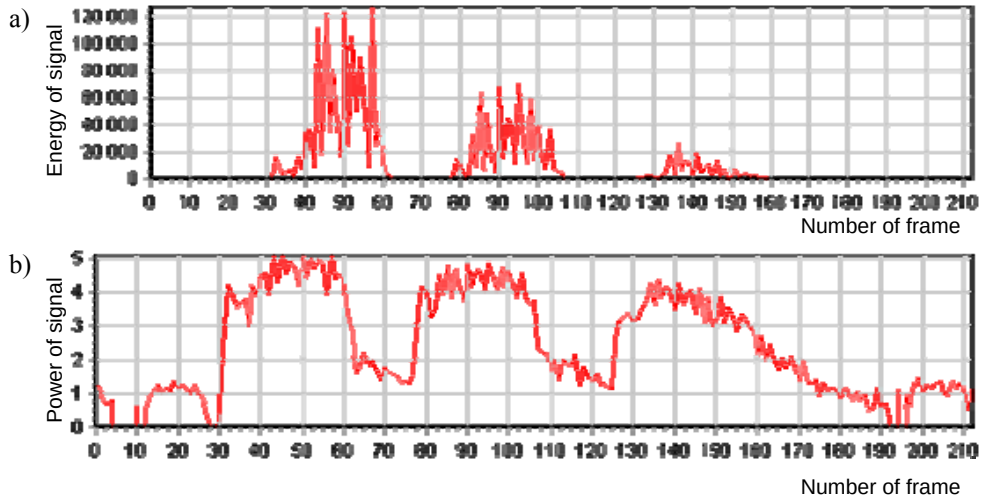

Fig.5 Energy (a) and power distribution (b) of the word "property"

6. THE FRAMES ENERGY AND FRAMES POWER NORMALIZATION

Before sound and silence detection, the energy and power of signal are normalized. After that, all values have the same range [0–1]:

$$E_{norm}(l) = \frac{E(l)}{\max\{E(1),\ldots,E(K)\}} \quad 1 \le l \le K \tag{8}$$

where:

l – number of current frame,
K – number of all frames,
$E(l)$ – energy of the l-th frame,
$E_{norm}(l)$ – normalised energy of the l-th frame.

The power of the signal P is normalized similarly:

$$P_{norm}(l) = \frac{P(l)}{\max\{P(1),\ldots,P(K)\}} \quad 1 \le l \le K \tag{9}$$

where:

$P(l)$ – power of the *l*-th frame,

$P_{norm}(l)$ – normalised power of the *l*-th frame.

Before normalisation amplitude of signals can be different in dependence on uttered word, for example for quietly or loudly uttered words. It was clearly depicted in Fig. 6.

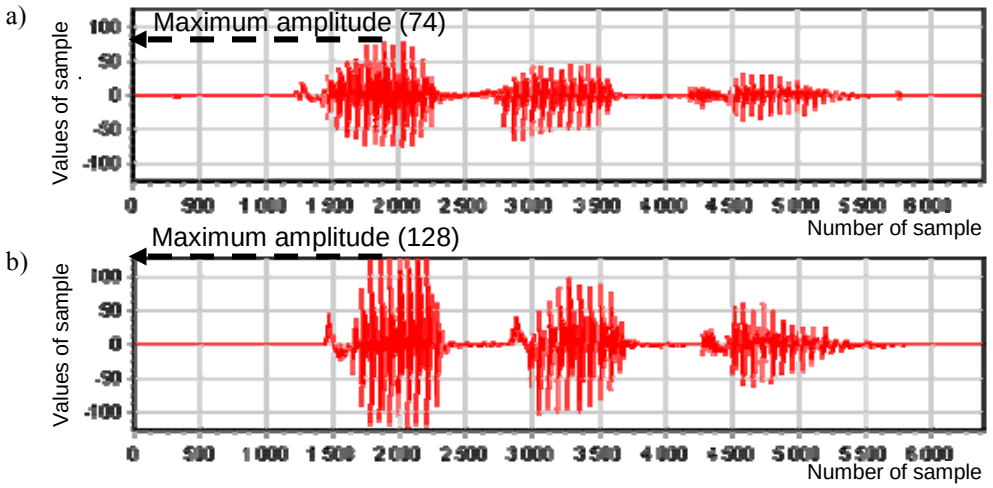

Fig.6 Differences between signal amplitudes of the word "property"
for quietly uttered word (a) and for loudly uttered word (b)

From Fig. 7 follows that normalization process does not change shape of signal and only values are proportionally changed.

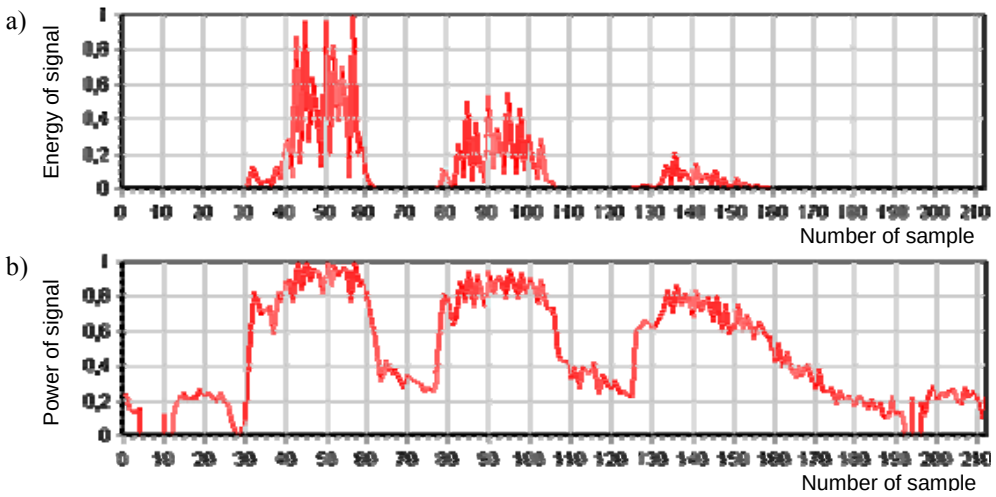

Fig.7 Energy (a) and power (b) of the signal for word "property" after normalization

7. ENERGY AND POWER LEVEL DETECTION

After normalization, real value of the amplitudes of signals is not important. For this reason, additional signal parameters can be introduced. These parameters determined the energy level threshold *Pe* and the power level threshold *Pm*. Mentioned parameters are used to detection of the lower and upper boundary of uttered words. Obviously, parameters *Pe* and *Pm* are also normalized. In the word detection process thresholds *Pe* and *Pm*, should be appropriate determined. Hence, uttered word is isolated between energy and power levels, where energy (power) is greater than values of the threshold. Information about the energy of the signal allows to determine frames with great amplitudes. Information about the power of signal allows to determine frames with low amplitudes. For this reason the both *Pe* and *Pm* parameters are always calculated. In the power chart can observe that signal values are expanded in zero-values areas – near silence (Fig. 8). It is not visible in the energy chart (Fig.9).

In the next stage number frames, where energy (power) level is greater than threshold level, is determined. In another step initial and final sample is searched. It was presented in the Fig. 8.

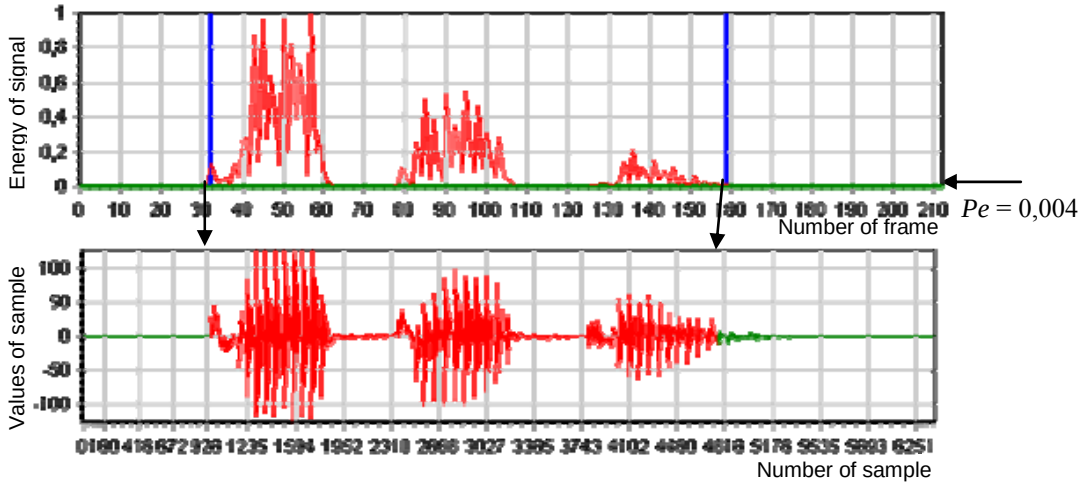

Fig.8 Boundaries of the word detection with energy level advantages. Red colour – detected single word

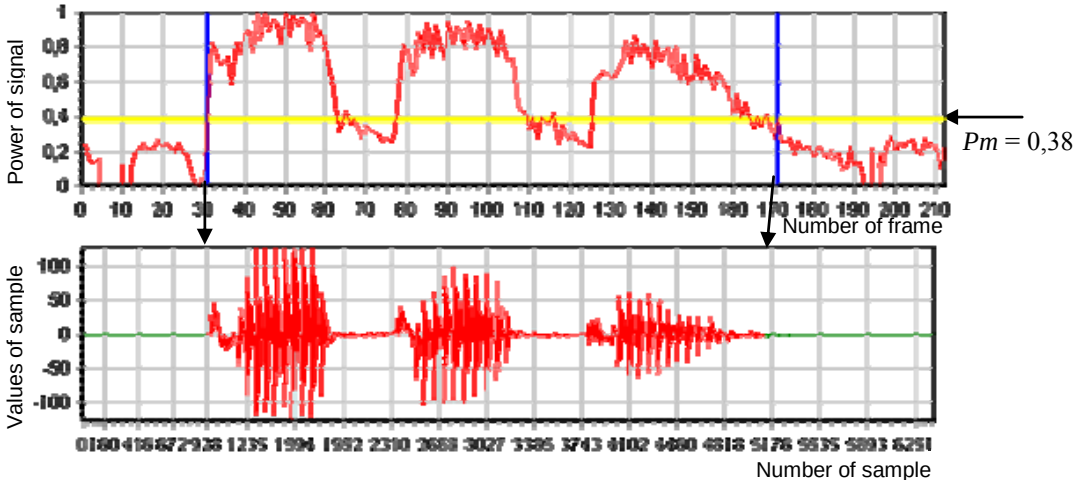

Fig.9 Boundaries of the word detection with power level advantages. Red colour – detected single word

After removing some frames, which lie outside of the interval, full word is formed (Fig. 10).

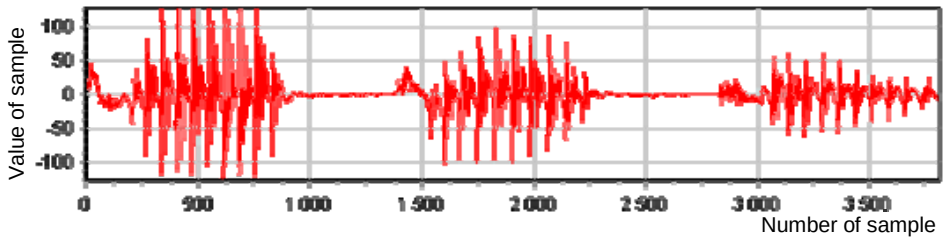

Fig.10 Results of isolation of the word „property”

From carried out investigation follows that energy (power) threshold level can be established differently but it does not significantly improve the word detection.

8. SELECTION OF THE PARAMETERS IN THE WORD DETECTION ALGORITHM

In the word detection algorithm, where the uttered word should be properly isolated, many parameters are selected:

- Sound threshold value – *Pd*,
- Time of silence before the uttered word,
- Time of silence after uttered word,
- Length of the frame (in milliseconds) – *D*,

- Power threshold level *Pe*,
- Energy threshold level *Pm*.

If the proposed system works in a clear environment, then extraction of the words is possible. For theses assumptions the parameters *Pe*=0,004 and *Pm*=0,38 were experimentally selected. In cases where a background noise will be too large, isolation and recognition of words can be carried out. From our investigation it follows that short, unexpected noises (phone bell, noisy scream, etc.) could not be effectively eliminated. On the other hand a constant noise appears in the room very often. The constant noises are for example music, working machines, fans, etc. Some of these noises can be eliminated by use of the directional microphones. In the first step silence and sound level should be determined. It can be computed by means of equations (1) or (2). In experiment carried out the best results of the word recognition have been obtained for constant *b*=4. In experiments popular chains of commands have been tested: „start, stop, close, open, save, reset, property, copy, paste and exchange". Word extraction is possible for the next restrictions: 12,5*ms* (100 samples) of silence before analyzed word and 100*ms* (800 samples) after end of the word. For precise word detection the second parameter is more important. In dependence on used parameters, the operator can utter from 1,5 to 1,8 commands per second (Fig. 11). It means that in our system average command detection time is inside the range of 556*ms* – 651*ms*.

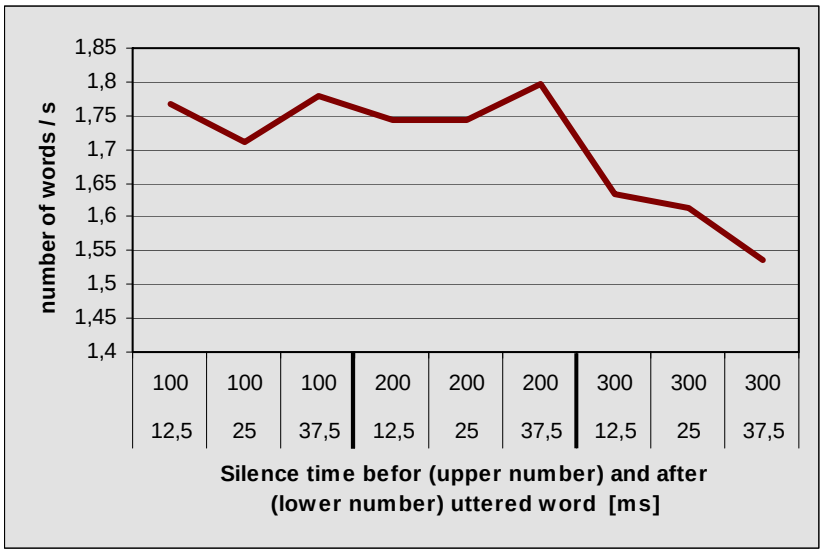

Fig.11 Promptness of the words processing

9. CONCLUSIONS

On the basis of recognised words, voice commands can be applied in computer programmes, for example in the MSAA Microsoft environment [7]. It follows from these investigations that word level recognition is satisfactory for most cases. People with disabilities are part of the population that benefit from using word recognition programs. It is especially useful for people with hand disabilities that require alternative input for support with accessing the computer.

BIBLIOGRAPHY

[1] ADAMCZYK A. WIŚNIEWSKI M., Segmentacja sygnałów mowy, Biuletyn Instytutu Automatyki i Robotyki WA, Nr. 13, 2000.
[2] ALWAN A., Wideband Speech Coding with Linear Predictive Coding (LPC). University of California. LA Department of Electrical Engineering, 2002.
[3] http://sound.eti.pg.gda.pl/student/eim/synteza/adamx/ - Efekty dźwiękowe, autor: Adam Michalski.
[4] KOVACEVIC M.A., EKEBERG J., DEHLBOM A., EKEBERG J., GARIAZZO G., LÄSTH E., TRONCOSO V., Ericsson T18s Voice Dialing Simulator, Report KTH-S3-2E1366 Stockholm, 2000.
[5] PIASECZKI M., ZOŚKO Sz., Rozpoznawanie granic słowa w systemie automatycznego rozpoznawania izolowanych słów, materiały konferencyjne KOSYR'99. Oficyna Wydawnicza PWr, 1999, str. 387-391.
[6] PORWIK P., BĘDKOWSKI K., ŻELECHOWSKI Ł., LISOWSKA A. Specialised LINUX Software aiding work of blind persons. Journal of Medical Informatics & Technologies. 2003, Vol. 5, pp. MT145–MT153.
[7] PORWIK P., Isolated word descriptors as control parameters of the computer applications. Journal of Medical Informatics&Technologies. 2006, Vol. 10, pp. 35–46.

[8] PORWIK P., SZCZEPANKIEWICZ M. The Voice Synthetiser of Polish Text for Blind Persons. Journal of Medical Informatics&Technologies. 2002, Vol.4, pp. MT101–MT109.
[9] SOZAŃSKI K., Połówkowo-pasmowe filtry cyfrowe. Zielona Góra: Red. Wyd. Naukowo-Technicznych UZ, 2002, str.213–218