

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800

Open access books available

122,000

International authors and editors

135M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Speech Recognition Based on the Grid Method and Image Similarity

Janusz Dulas
Opole University of Technology
Poland

1. Introduction

The problem of communication between a man and a machine is very old. First constructors had to decide how to transmit information from a man to machine and vice versa. This problem still exists and each engineer who designs a new device must decide how the communication between operator and a machine will be done. Simple devices use buttons and Light Emit Diodes [LED], more complicated – keyboards and screens. Fast technical development and numerous scientific research allowed to use also voice for this purpose. Here there are two different problems: voice producing and voice recognition. The first one is not very difficult, in the simplest case the machine could record a set of words which would be used for communication. Nowadays there are specialised integrated circuits (speech processors) which enable recording and reproducing whole words and sentences. The second problem – voice recognition is more complicated. First of all people are different and say the same words in a different way. Secondly, the way of speaking depends on many aspects like health of the speaker, his mood or emotion. Thirdly, we are living in noisy environment so usually together with speech signal we also obtain different noises. There are a lot of different methods used for automatic speech recognition. The most popular is HMM – Hidden Markov Model (Junho & Hanseok, 2006; Wydra, 2007; Kumar & Sreenivas 2005; Ketabdar at al., 2005), which uses sequences of events (states), where the probability of being in each state and the probability of transition to the other states are counted. Each state is described by many, mostly spectral parameters. There are also other, less popular methods used for automatic speech recognition like Neural Network Method (Vali at al., 2006; Holmberg at al., 2005; Togneri & Deng, 2007), Audio-visual Method (Seymour at al., 2007; Hueber at al., 2007) and many others (Nishida et al., 2005). Nowadays there is a possibility to achieve more than 90% accuracy in automatic speech recognition. HMM method, although the most popular, is very complicated. Each state is described by the matrix with many spectral, cepstral and linear prediction parameters. It causes the need for many analyses and calculations during the automatic recognition process. In this chapter the author shows a new approach to this problem. The new method described here is simpler and faster than HMM method and gives similar or better results in speech recognition accuracy. Although it was tested in Polish, the rules can be adopted to other languages.

2. Some interesting voice signal properties

Usually the voice signal which will be analysed is first converted into the electric signal by the microphone. The most popular are dynamic and condenser microphones. No matter which one is used, we obtain a signal with frequency of between dozen or so hertz and some kilohertz. Figure 1 shows time characteristic for word "zero" spoken by the man.

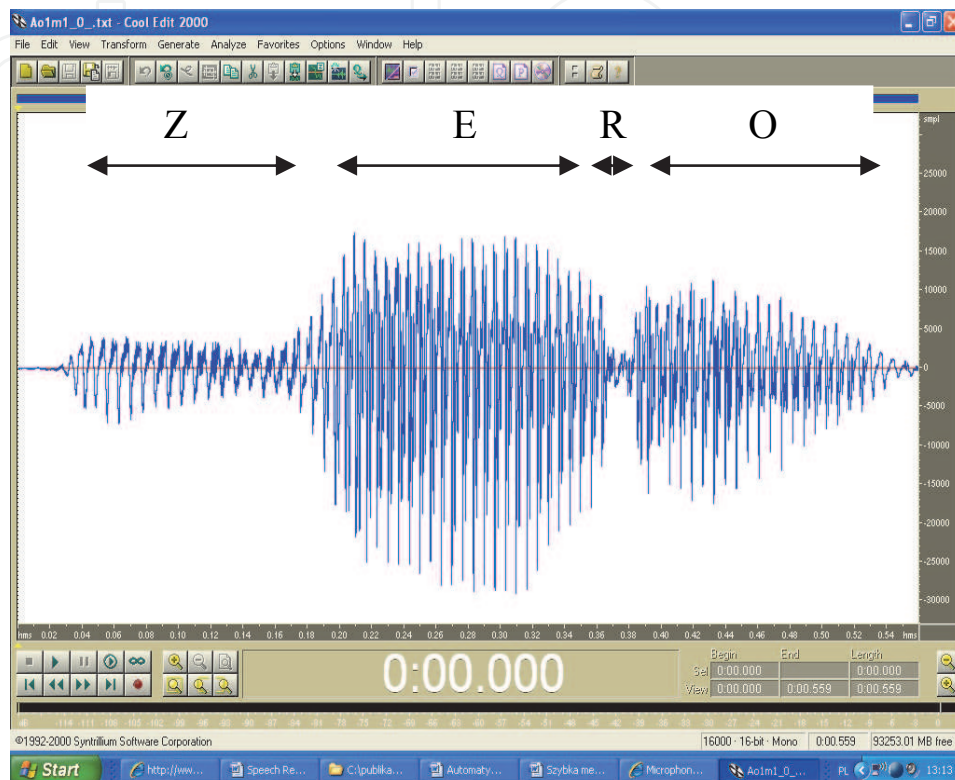


Fig. 1. The electric signal time characteristic for word "zero"

As is well known, each word could be divided into the smallest parts called phonemes. In Polish there are 37 phonemes which can built 95% of all words. Usually they are well visible in time characteristic. In figure 1 it is easy to find four phonemes for example by listening tests of different parts of this record. As is easy to observe, the phonemes could have different duration. For example in figure 1, the "R" phoneme has the smallest duration. They could also have different amplitudes. In figure 1 phonemes "E" and "O" have greater amplitudes than phonemes "Z" and "R". As the theory said - phonemes could be voiced and unvoiced. All vowels and some consonants are voiced. Voiced phonemes are easy to find because they include many repeatable basic periods inside of them. These periods have similar duration (between 2 and 10 ms) which is dependent on the speaker's sex and age. Men have bigger basic periods duration (usually 7...10ms) than women and children (usually 2.5ms). The shape of the signal in neighbouring basic periods is similar. Let's look at the magnifying part of the phoneme's "O" time characteristic (figure 2).

There are many basic periods with similar signal's shape and similar duration (about 7ms). Because the duration equals 7ms, it means this is the man's voice record. Another interesting voice feature is the envelope shape. In figure 3 there are shown three time characteristics of the word "zero" spoken by the man, woman and child.

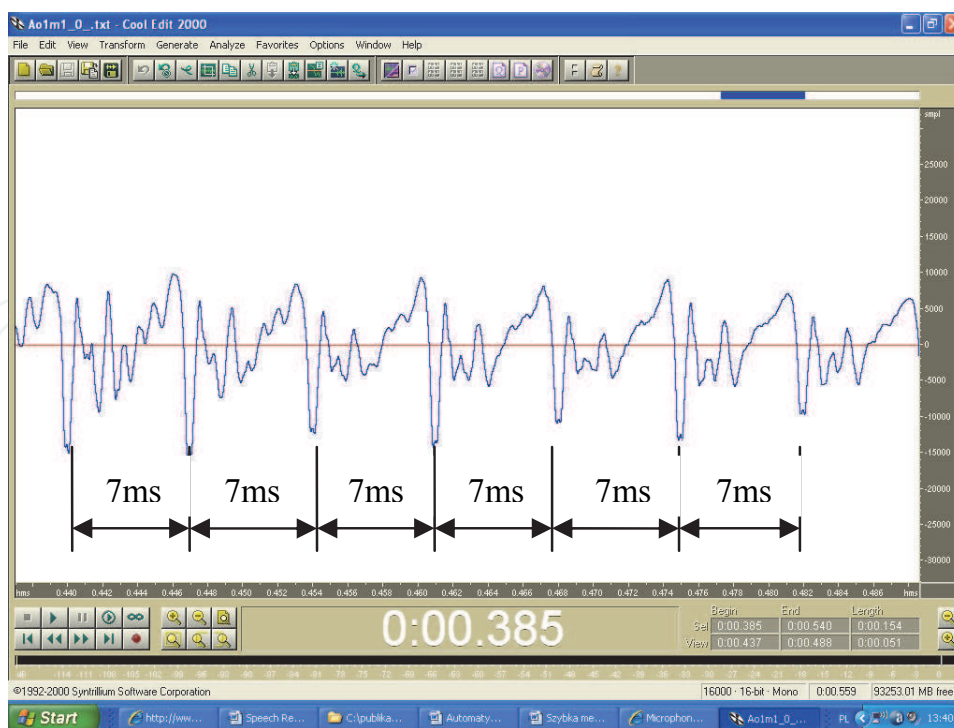


Fig. 2. Magnifying part of phoneme “O” time characteristic.

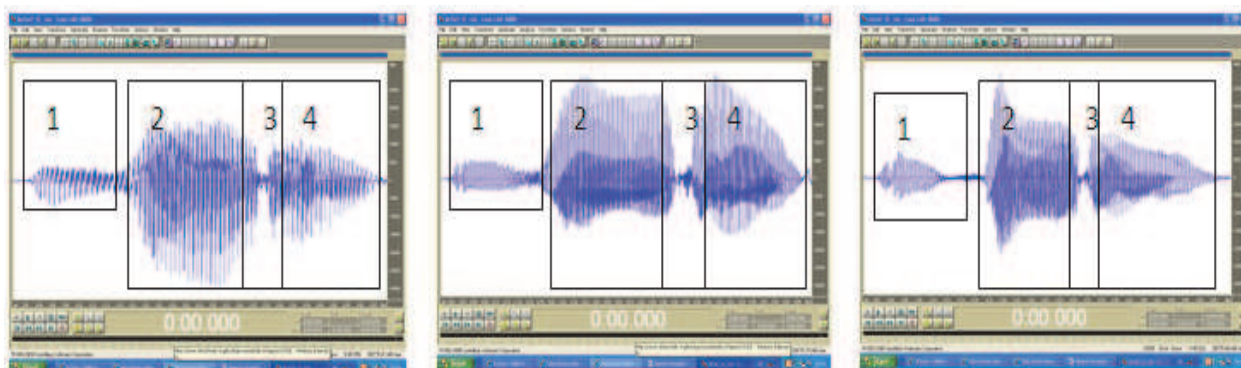


Fig. 3. The word “zero” spoken by the man, woman and child

In rectangles there are included four different parts of the records. The first has rather small amplitude, the second’s amplitude is bigger, the third’s smaller and the fourth’s bigger. Although these records come from different speakers, the same amount of different parts is included inside them. Maximum duration of each part could vary and is difficult to find but there is a possibility of finding the minimum duration what has been proven in authors research.

Another voice feature is fluently changing the signal shape between neighbouring phonemes. It is not just one point but the zone, which possible duration of many milliseconds.

In figure 4 there is shown such a zone between phonemes “z” and “e” in word “zero”.

As is shown in figure 4 the transient zone duration in this case equals about 30ms.

Another group of phonemes are noisy phonemes. These phonemes don’t include basic periods in their time characteristics but many irregular signals. In Polish there are some such phonemes. The example of noisy phoneme time characteristic is shown in figure 5.

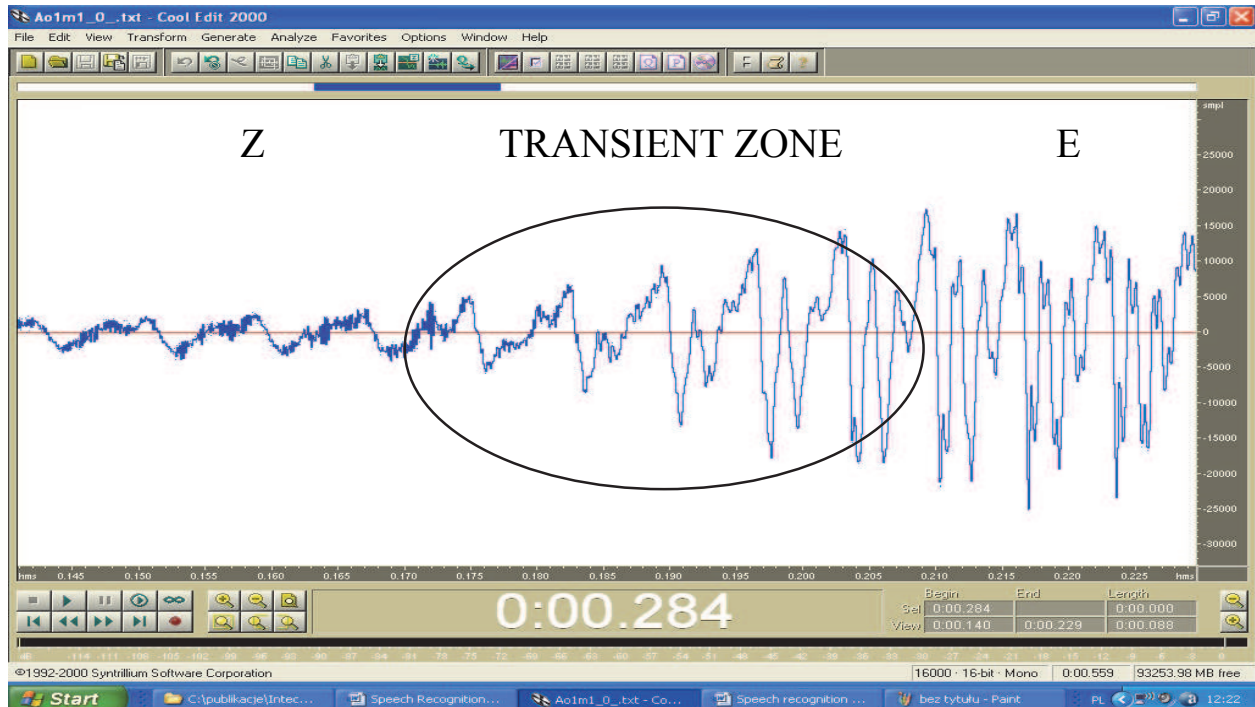


Fig. 4. Transient zone between phonemes “z” and “e” in word “zero”

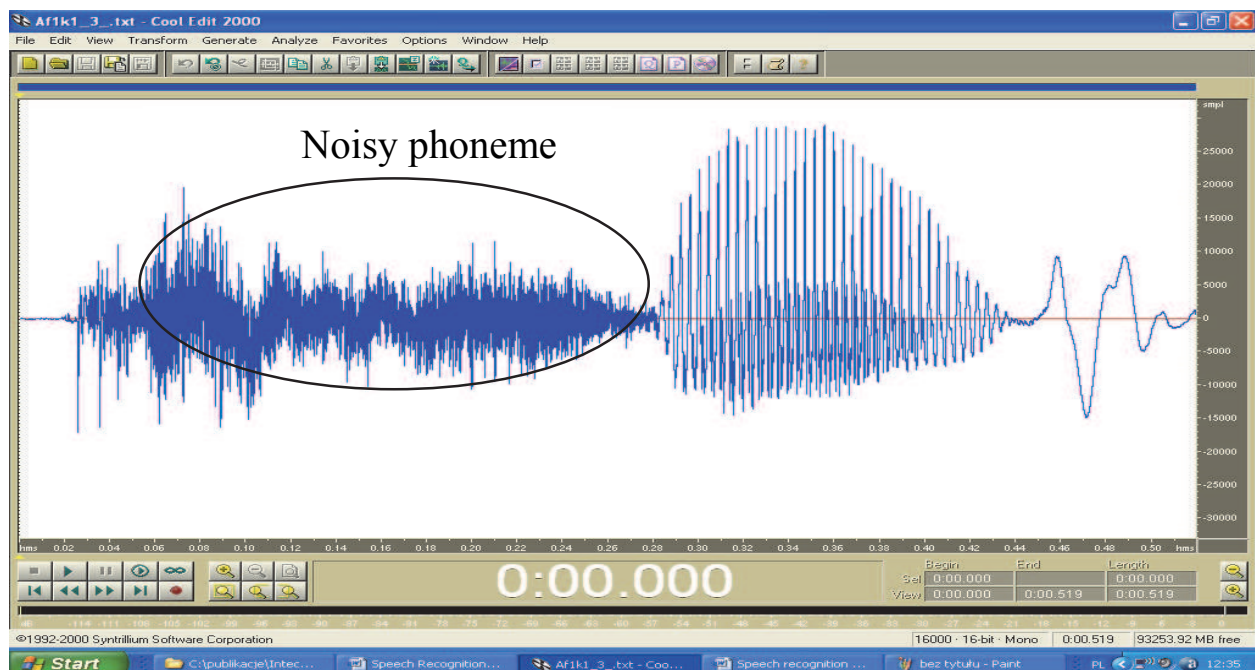


Fig. 5. The word spoken in Polish including a noisy phoneme.

The magnifying part of noisy phoneme from figure 5 is shown in figure 6.

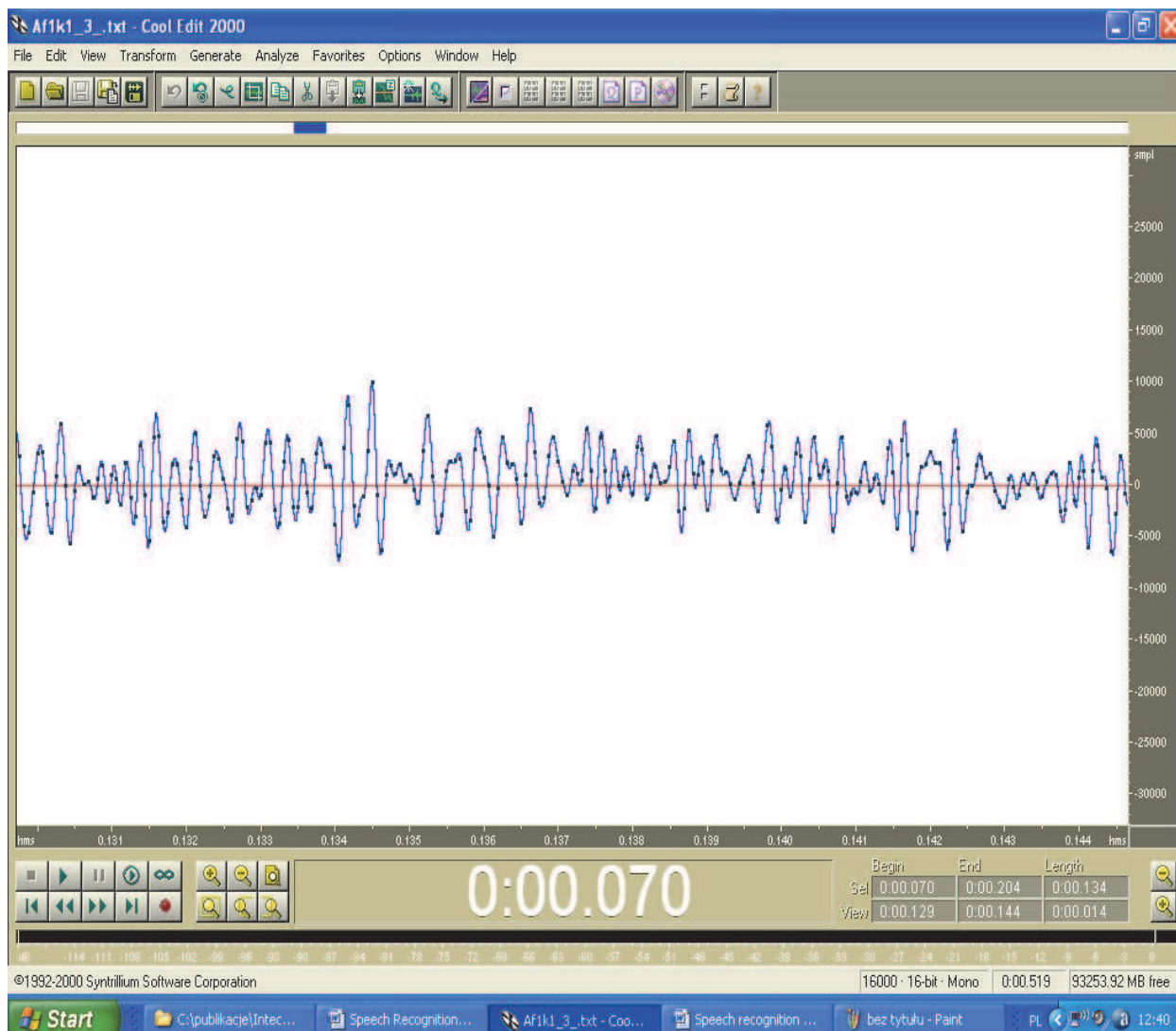


Fig. 6. A magnifying part of noisy phoneme time characteristic

As is easy to notice here there are no repeatable similar basic periods. The signal changes its amplitude very often. Sometimes the stronger signals appear but the moments of these events are accidental.

3. Envelope similarity analyses

As was mentioned in section 2, the envelopes of the same word spoken by different speakers have similar shapes. This feature will be used for automatic speech recognition. There are a lot of possibilities of the envelope shape describing. In author's research, each word (digits from 0 to 9 names spoken in Polish) was described by number of unique parts, their minimum durations and amplitude range. This operation was made on the 500 records set from speakers different sex and age. As results showed, this approach allowed to build envelope patterns of all digits names for all speakers. They are shown in table 1.

As was shown in table 1 each digit has its own envelope pattern with a defined number of unique parts (column 3), their minimum durations (column 4) and amplitude ranges (column 5). The amplitude values are defined in per cents of the maximum signal's value for

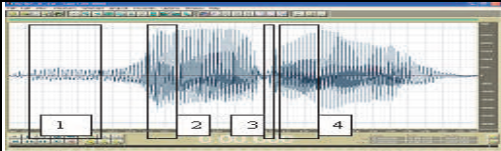
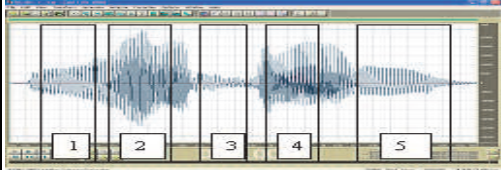
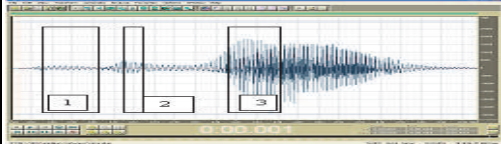
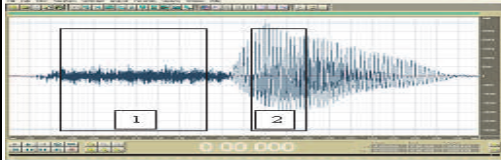
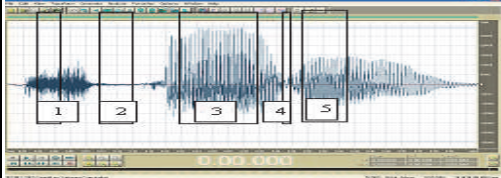
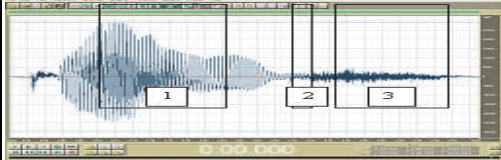
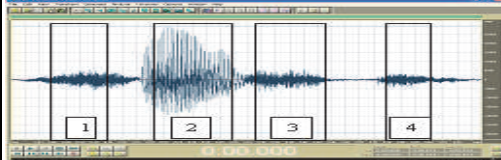
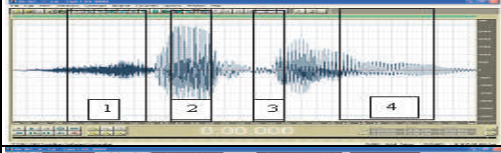
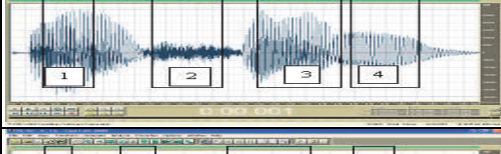
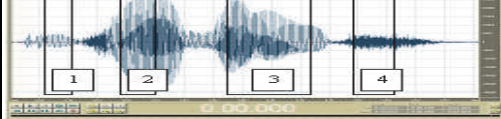
Digit	Envelope's shape	Number of parts	Minimum durations	Amplitude ranges
0		4	1.80ms 2.30ms 3.10ms 4.50ms	1.A<50% 2.A>53% 3.5-64% 4.A>31%
1		5	1.60ms 2.70ms 3-50ms 4-60ms 5-100ms	1.A<62% 2.A>44% 3-1%<A<26% 4-A>25% 5-A<50%
2		3	1.50ms 2.20ms 3.50ms	1.A<30% 2.A>29% 3.A>43%
3		2	1.130ms 2.40ms	1.6%<A<32% 2.A>68%
4		5	1.30ms 2.40ms 3.100ms 4.10ms 5.60ms	1.A>1% 2.A<4% 3.A>40% 4.3%<A<35% 5.A>13%
5		3	1.140ms 2.20ms 3.110ms	1.A>29% 2.A<23% 3.A>0%
6		4	1.80ms 2.100ms 3.100ms 4.70ms	1.2%<A<35% 2.A>26% 3.1%<A<56% 4.A>1%
7		4	1.100ms 2.50ms 3.30ms 4.110ms	1.3%<A<65% 2.A>36% 3.2%<A<28% 4.6%<A<59%
8		4	1.70ms 2.90ms 3.100ms 4.90ms	1.A>46% 2.6%<A<61% 3.16%<A<92% 4.5%<A<43%
9		4	1.40ms 2.70ms 3.150ms 4.80ms	1.1%<A<45% 2.A>46% 3.A>9% 4.A<50%

Table 1. Envelope patterns together with unique parts parameters

the whole words. As is easy to observe, the number of unique parts is different for different digits and their values are between 2 and 5 (for Polish). These patterns were tested on 500 records and all of them are compatible with them.

4. The grid method

The word's phonemes structure could be very important for automatic speech recognition. For digits' recognition it could be simplified to recognition of voiced phonemes and noisy phonemes. Table 2 shows simplified phonemes structure for digits 0-9 spoken in Polish.

Digit	Simplified phonemes structure (V-voiced phoneme, N-noisy phoneme)
0	V+V+V
1	V+V+V+V
2	V+V+V
3	N+V
4	N+V+V
5	V+V+N
6	N+V+N
7	N+V+V+V+V
8	V+N+V+V
9	V+V+V+V+V+N

Table 2. Simplified phonemes structure for digits 0-9 spoken in Polish

According to table 2 some digits include noisy phonemes (3, 4, 5, 6, 7, 8, 9) and some don't (0, 1, 2). Also the number of voiced phonemes is different (from 1 to 5). For finding voiced phonemes the author's Grid method was used.

4.1 The basic periods finding

The first step according to the grid method is finding basic periods included in voiced phonemes. At the beginning the algorithm tests 100ms signal before the beginning of the recorded word. This way the mean value of the noises are computed and "zero" level of the signal is found. Then, all samples of the signal are tested and local minimums are found. Three criteria are used here:

1. The sample is a minimum if the next sample's value and the previous sample's value is greater than the tested sample.
2. The time between the previous sample and tested sample must be greater than 2ms
3. The tested sample's value must be lower than the "zero" level.

This way, the local minimums are found and written to the matrix as a number of samples in which they were found.

Figure 7 shows the result of this operation. The local minimums were matched by the circles.

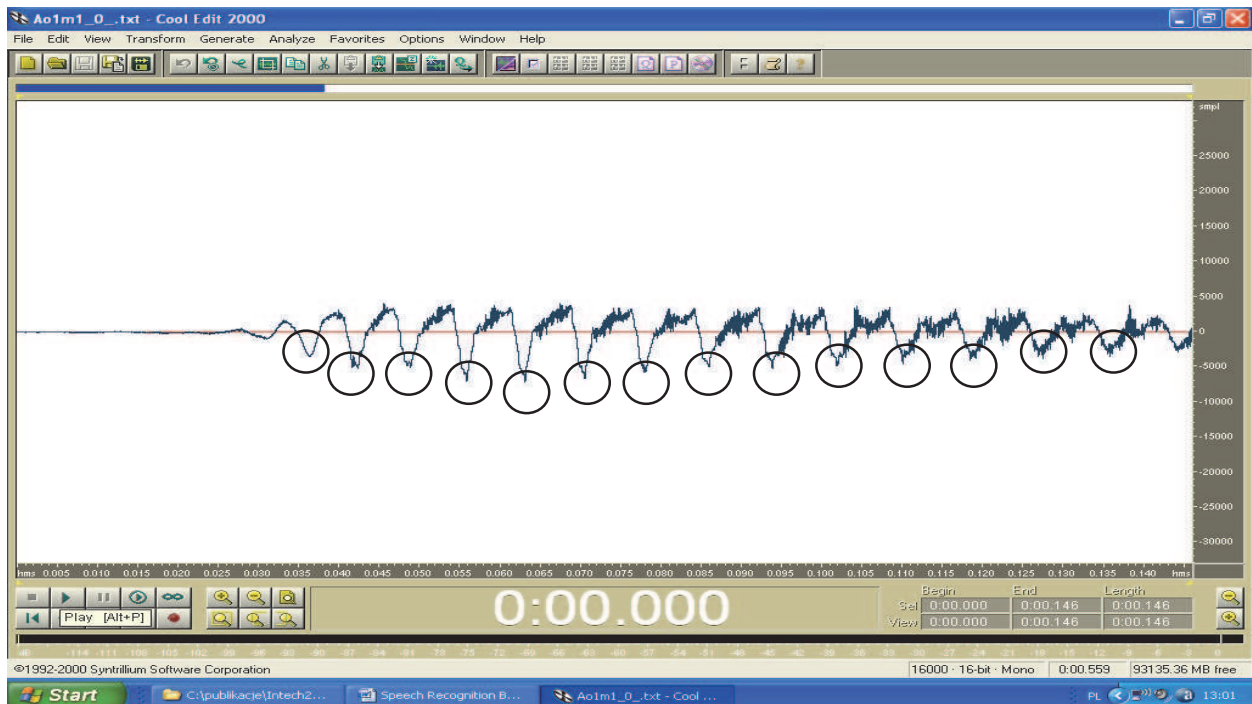


Fig. 7. Local minimums found in the speech signal

4.2 The grids fitting

The next step is coding each basic period in the matrix. This is made by putting grids on them. The rule is showed in figure 8. In cells of the grid "1's" or "0's" are automatically written. Ones - if there is a signal inside a cell, zeros - if there is no signal.

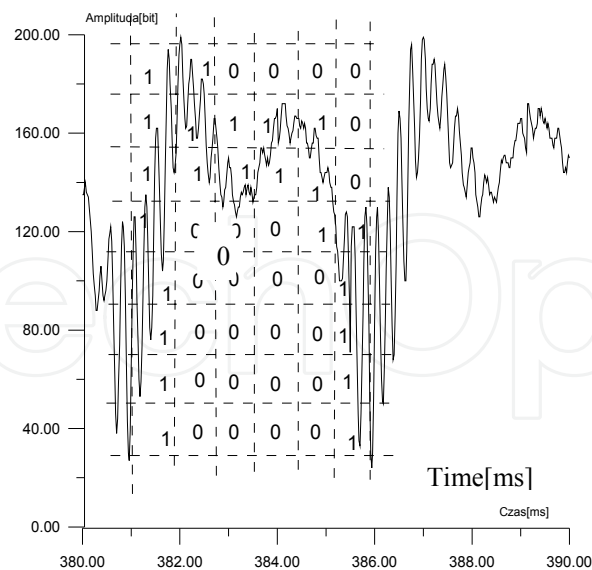


Fig. 8. The basic period coding

The size of each grid is automatically fitted into the signal. The highest - to the signal's amplitude and the widest - to the basic period duration. This way the shape of the signal is coded independently of the personal speaker's voice feature. If the amplitude or duration

change the size of each cell is changes proportionally. Finally, the binary matrix is obtained. Fig.9 shows this matrix for signal from Fig.8.

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 9. The binary matrix obtained from the grid

4.3 The similarity coefficients computing

For one word there is a possibility to obtain even some hundred binary matrixes. This amount depends on the phonemes' number in this word and the speaker's sex. For the same length of the word women and children records have more grids than the men. This is because of the less basic periods duration for women and children voices. After the binary matrixes obtaining, the similarity among them is computed. Each matrix is automatically compared with the five matrixes before it and the five matrixes after it (all bits on the same position of matrixes are compared). If more than 88% bits are the same, the grid is matched as "similar" if not as "not similar". This way, all the grids get their similarity coefficients. This coefficients could have value between 1 (there is no similar grids around it, so it is only one such matrix) and 11 (there are 5 similar matrixes before and after tested matrix plus this matrix).As author's research showed, if there 1 or more similarity coefficients one after another, with value higher than 27% of the maximum similarity coefficient in the word, which last (together) at least 10ms it means that there is a voiced phoneme. The lower number of such coefficients means that there is no phoneme (break) , unvoiced phoneme or noisy phoneme. It could also show the zone where one voiced phoneme is finishing and another starting. This is a very important observation because it allows to find all voiced phonemes inside each word. This parameter (number o phonemes) will be used in the automatic speech recognition process. The same research showed that the brake between phonemes insists if three conditions are performed simultaneously:

1. Duration of grids coming one after another, with similarity coefficient lower than 27% of the maximum similarity coefficient in the word, must be greater than 14ms.
2. Time between the previous break and tested grid must be greater than 88ms
3. Time between the beginning of the last voiced phoneme and tested grid must be greater than 42ms

The example of the similarity coefficients for word "zero" is shown in table 3.

As is easy to observe in table 3, grids from 7 to 20 have similarity coefficients greater than 2, so there is a voiced phoneme. Then there are grids with smaller coefficients, although sometimes there is a grid with the bigger one. As was mentioned earlier - one or more grids lasted more than 10ms (in this case at least 3 grids) with coefficients greater than 2 ($27\% \times 11 = 2,97$) depict the voiced phoneme, so one or two such a grid means nothing. This is the part of the record where the one voiced phoneme (z) changing to another (e), so this is the transient zone. The grids from 40 to 82 have again greater similarity coefficients so it means that another voiced phoneme (e) was found . Grids from 83 to 89 have small

Number of grid	Similarity coefficients	Phoneme	Number of grid	Similarity coefficients	Phoneme	Number of grid	Similarity coefficients	Phoneme
1	1		39	2		77	7	E (continuation)
2	1		40	5		78	11	
3	3		41	4		79	11	
4	1		42	3		80	10	
5	2		43	4		81	9	
6	2		44	3		82	8	
7	5	Z	45	5	E	83	2	
8	8		46	4		84	3	
9	7		47	6		85	1	
10	8		48	5		86	2	
11	11		49	6		87	1	
12	10		50	4		88	1	
13	10		51	7		89	1	
14	9		52	8		90	3	
15	9		53	8		91	5	
16	7		54	10		92	6	
17	7		55	11		93	5	
18	3		56	11		94	6	
19	3		57	11		95	6	
20	7		58	9		96	5	
21	2		59	9		97	6	
22	4		60	10		98	6	
23	2		61	11		99	4	
24	1		62	11		100	6	
25	2	63	11	101	8			
26	3	64	11	102	7			
27	2	65	11	103	7			
28	1	66	11	104	8			
29	2	67	11	105	9			
30	1	68	10	106	11			
31	1	69	9	107	11			
32	1	70	11	108	11			
33	1	71	11	109	11			
34	3	72	11	110	11			
35	2	73	11	111	11			
36	3	74	11	112	6			
37	6	75	11	113	7			
38	1	76	11		

Table 3. Similarity coefficient in word "zero" and found voiced phonemes

coefficients, but this time it is because the unvoiced phoneme “r” is found. Grids from 90 to 113 have greater coefficients what means that another voiced phoneme (o) was found. This way, using only similarity coefficients, three voiced phonemes were found. This method was used for finding a number of voiced phonemes for digit’s from 0 to 9 records (50 records for each digit = 500 records). Another example for digit “1” (in Polish “JEDEN”) is shown in table 4.

Number of grid	Similarity coefficients	Phoneme	Number of grid	Similarity coefficients	Phoneme	Number of grid	Similarity coefficients	Phoneme
1	4	J	25	2	E Continu- ation	49	8	E-N Continu- ation
2	4		26	5		50	8	
3	4		27	8		51	5	
4	7		28	8		52	7	
5	5		29	7		53	6	
6	6		30	6		54	7	
7	2		31	6		55	7	
8	1	32	1	56	7			
9	1	33	1	57	6			
10	1	34	1	58	6			
11	2	35	1	59	1			
12	1	36	1	60	5			
13	3	37	2	61	7			
14	2	38	1	62	5			
15	2	39	2	63	7			
16	2	40	1	64	8			
17	4	41	1	65	10	E-N Continu- ation		
18	3	42	1	66	10			
19	6	43	2	67	8			
20	1	44	1	68	7			
21	5	45	4	69	6			
22	8	46	4	70	6			
23	8	47	7	71	1			
24	10	48	7			

Table 4. Similarity coefficient in word “jeden” and found voiced phonemes

Here 4 phonemes were recognized (“D’ phoneme hadn’t basic periods). Two last phonemes (E-N) are not divided by the small coefficients so there is a transient zone. As research shown such a zones are present if the phoneme last at least 100ms and inside of it there are grids with similarity coefficients less than 70% of the maximum similarity in the word, and

before, and behind them there are grids with higher than 70% similarity coefficients. Another example of digit "8" (in Polish "OSIEM") is shown in table 5.

In this case (tab.5) there is one noisy phoneme between voiced phoneme "O" and "E" and one break between phonemes "E" and "M". This break is only 15ms long but it is enough (according to the rules mentioned above the break exist if the duration of the grids with less than 27% of the maximum similarity coefficients is bigger than 14ms). This record was made by the man whose basic periods equal 7,5ms. More details about noisy phoneme finding is placed in section 5.

Number of grid	Similarity coefficients	Phoneme	Number of grid	Similarity coefficients	Phoneme	Number of grid	Similarity coefficients	Phoneme
1	3		25	4	Noisy Phoneme (continuation)	49	7	
2	3		26	4		50	3	
3	2		27	2		51	1	Break
4	2		28	1		52	1	M
5	3	O	29	6		53	3	
6	5		30	2		54	4	
7	5		31	3		55	6	
8	5		32	1		56	2	
9	4		33	2		57	7	
10	3		34	2		58	8	
11	1		35	2		59	4	
12	1		36	1		60	10	
13	2		37	3		61	7	
14	3		38	2		62	8	
15	3	39	2	63		6		
16	5	40	6	64		5		
17	2	41	6	65		4		
18	2	Noisy Phoneme	42	6		66	3	
19	2		43	3	67	4		
20	4		44	4	68	6		
21	2		45	5	69	3		
22	5		46	6	70	3		
23	5		47	7	71	2		
24	5		48	8		

Table 5. Similarity coefficient in word "jeden" and found voiced phonemes

5. A noisy phoneme finding

As was mentioned in point 4, digits from 3 to 9 spoken in Polish include noisy phonemes. In figure 10 there is a time characteristic for digit "3" spoken in Polish by the woman.

As the author's research showed the duration of the noisy phoneme is usually bigger than 50ms, there are no basic periods inside them and they could be placed in different parts of the word once or more times. Figures 11-13 show different possibilities.

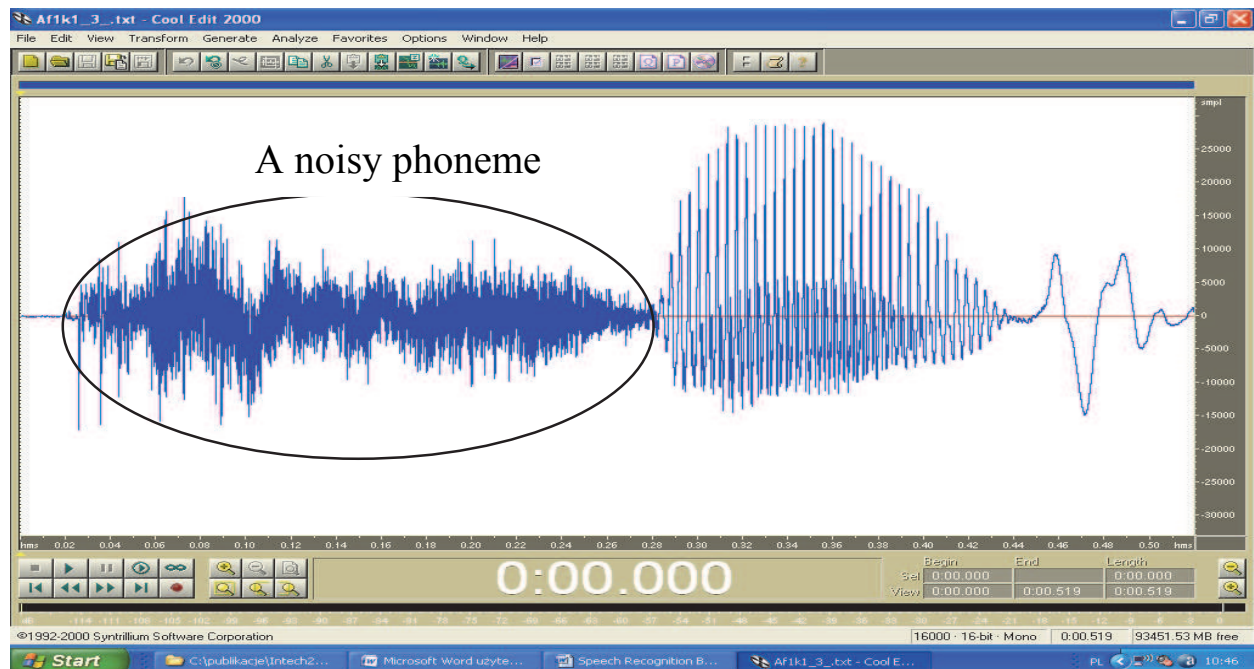


Fig. 10. Digit's three name spoken in Polish be the woman

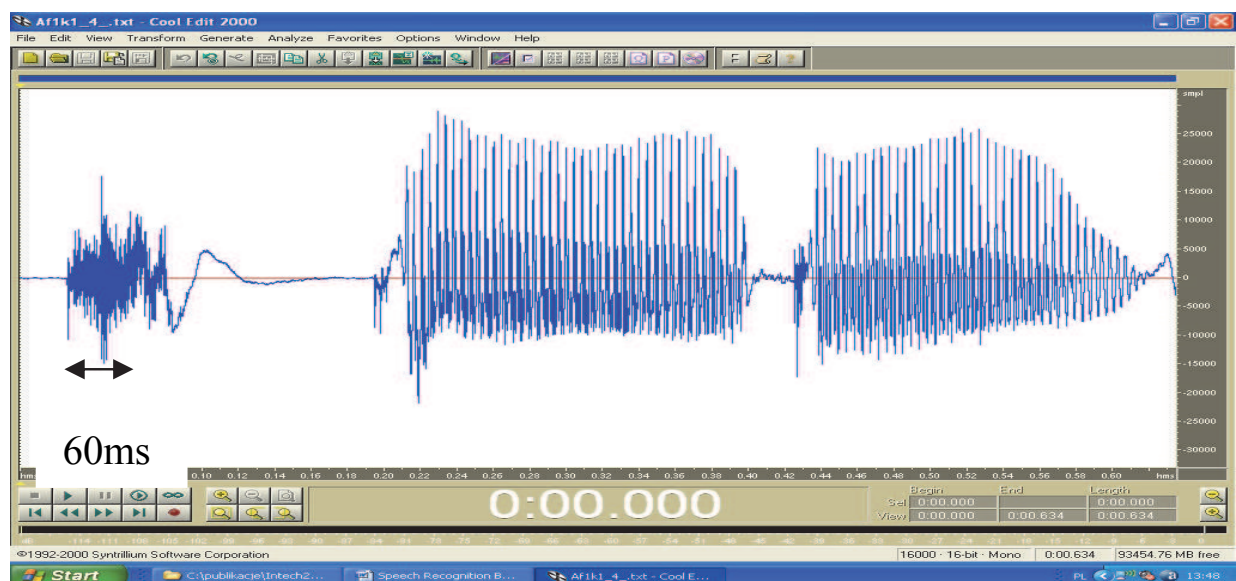


Fig. 11. The word with e noisy phoneme placed at the beginning of the word (digit 4 spoken in Polish)

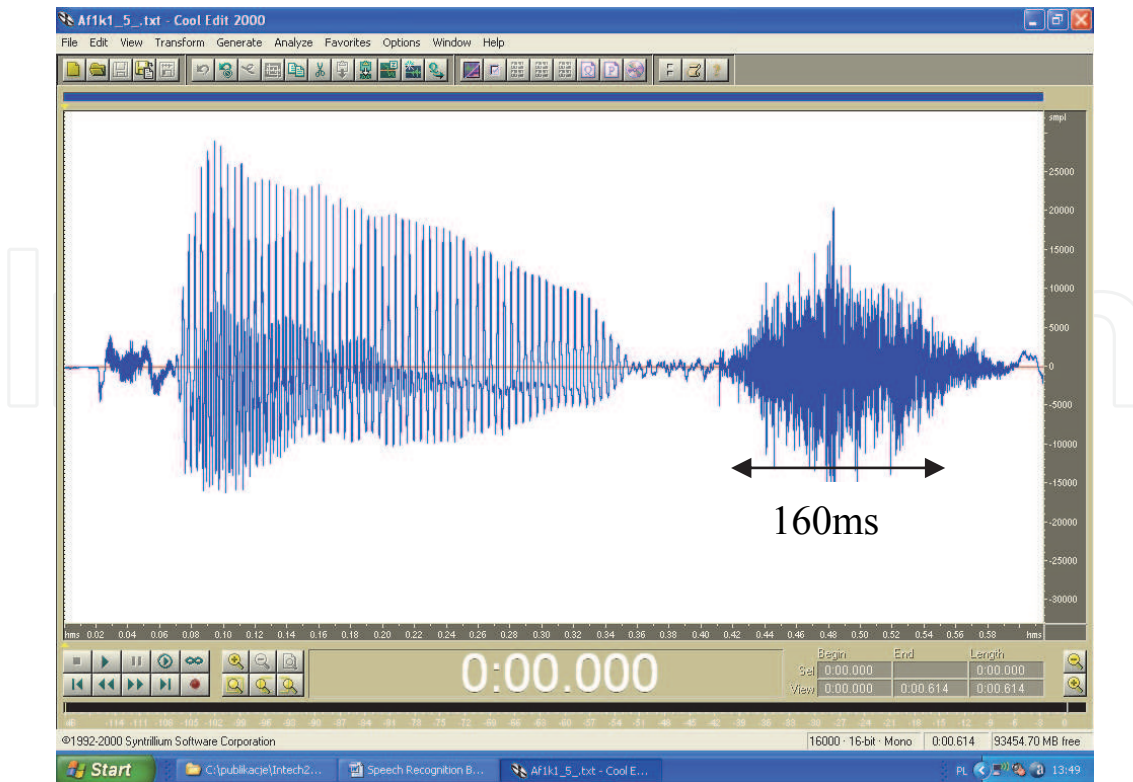


Fig. 12. The word with noisy phoneme placed at the end of the word (digit 5 spoken in Polish)

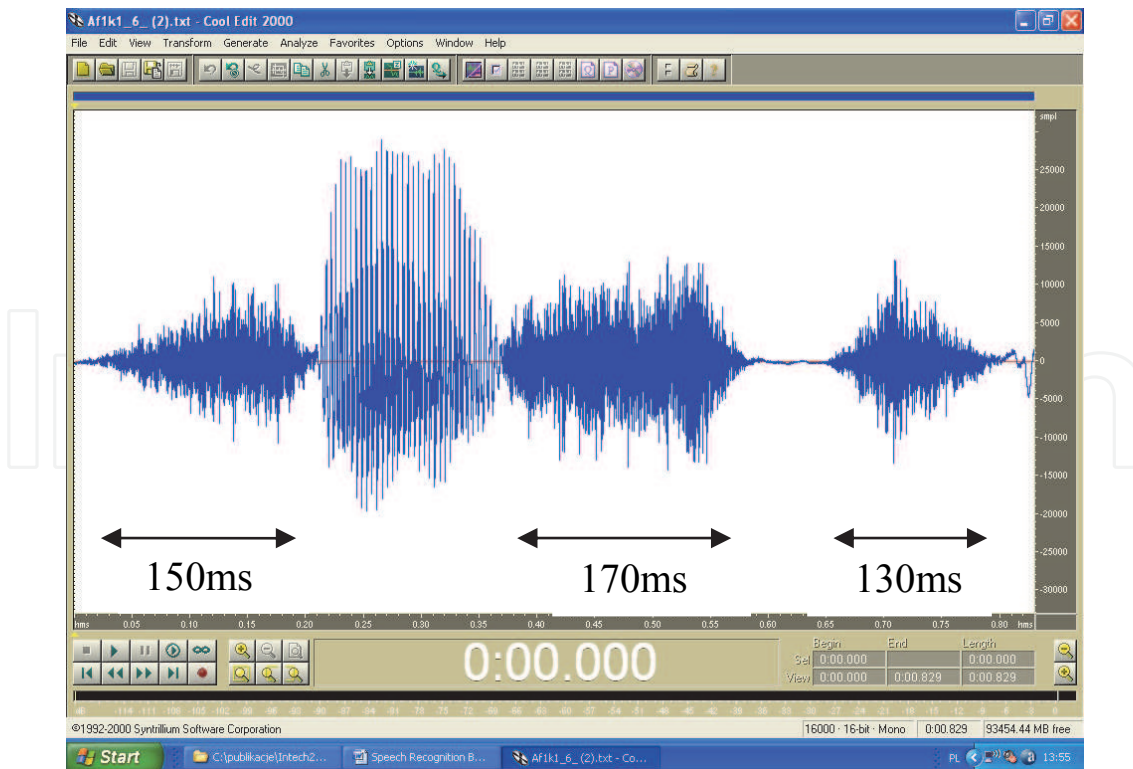


Fig. 13. The word with a noisy phonemes placed at the beginning and at the end of the word (digit 6 spoken in Polish)

Sometimes the noisy phoneme has a very small amplitude that could be mistaken with outside noises. (Fig.14)

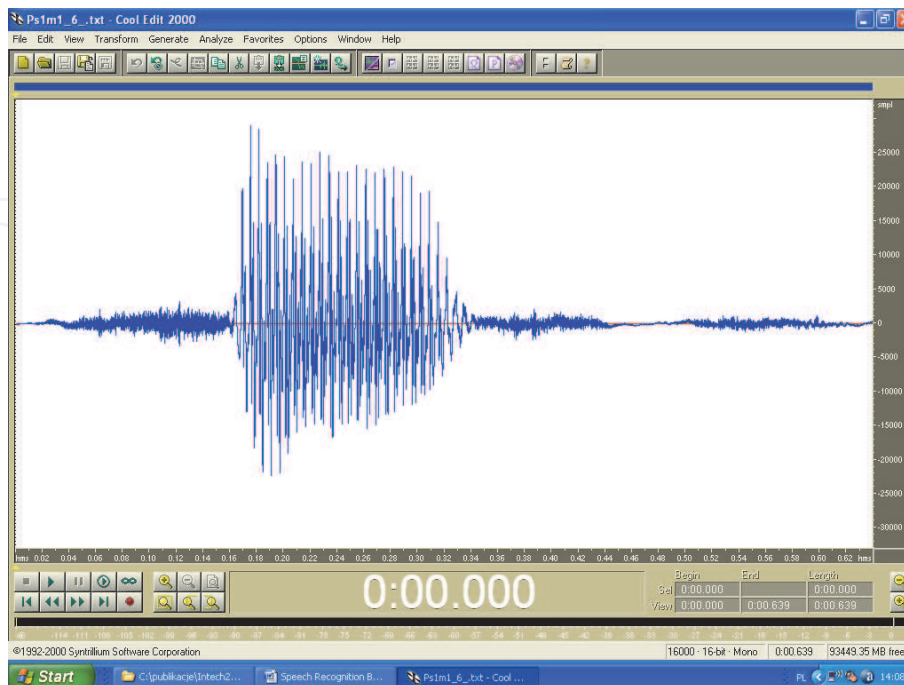


Fig. 14. A word with small amplitude noisy phonemes.

Another noisy phoneme's interesting feature is a big number of local extremes placed in small distances one after another. Figure 15 shows an enlarged part of the noisy phoneme's time characteristic.

As is easy to observe (in fig.15) The time between neighbouring extremes is usually less than 0,5ms.

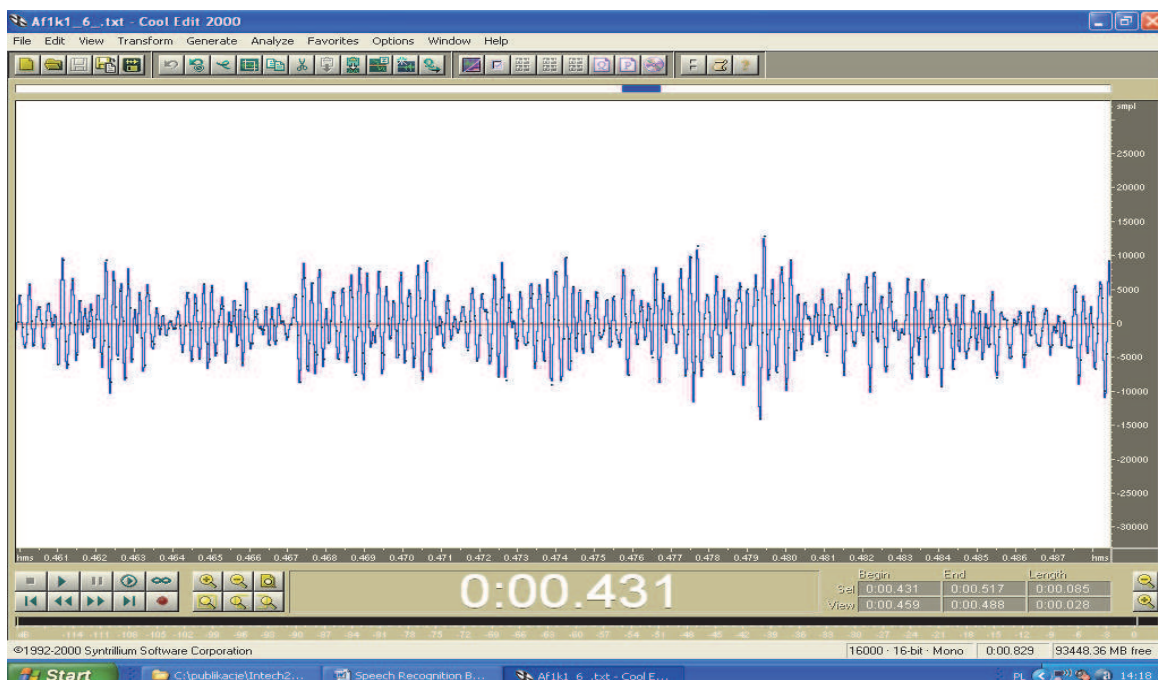


Fig. 15. A big number of local extremes inside the noisy phoneme.

6. The automatic digit's names recognition algorithm

All features described in the previous units were used for constructing an automatic digit's name recognition algorithm. It was implemented in Delphi software and automatically computed all necessary parameters. The complete algorithm is showed in figure 16. At the beginning the file with the recorded word should be loaded. Then the signal's "zero level" is found. Next, the local minimums are found and borders of the basic periods are computed. Then, the grids with 5×7 cells are automatically fitted to the basic periods. From here the binary matrixes are obtained. Next, the similarity coefficients are computed. According to them the number of the voiced and noisy phonemes is calculated. The next step is constructing of the word's phoneme structure. Here the voiced and noisy phonemes are placed in the right order, depending on their starting and ending time.

The received word's phoneme structure is compared with 9 structures (number of voiced phonemes for digit "0" and "2" for Polish is the same). If no one of them is the same, the word is treated as "unrecognizable" and the algorithm finishes its work. If the word's structure is the same as one of the 9-s shown in the table, the respective envelope pattern is compared with the recorded signal. Now there are two possibilities. If the envelope pattern agrees with the signal envelope the algorithm gives as a result the right digit, if not the word is treated as an unrecognized. As is easy to observe the envelope analyses is more important for digits "0" and "2" than for the others because of their the same phoneme's structure. For other digits this analyses result makes the recognition process more exact. The rules presented in the algorithm could be modified. For example instead of treating the word as "unrecognized" the algorithm could calculate the percentage similarity to all possible structures. The same could be made for the envelope analyses. Also here the similarity calculations are possible. As a result the algorithm always could find the solution but the user should be informed if the result is reliable or percentage calculated.

7. The way of research making

The envelope analyses was made in some steps. First, the time characteristics for each digit and different speakers was reviewed. From here the number of units was obtained. Then each unit was described by a minimum duration and amplitude range. Next, for another record these two values were tested and corrected in order to achieve properly the units number. This method is shown in figure 17.

In the similar way the number of voiced and unvoiced phonemes was found and corrected. After each change which improve the recognition results for any digit the previous digits were tested. This way the algorithm was tested for each digit. About 30% of records weren't used for parameter correction because their recognition results were proper from the beginning, also sometimes any correction in one parameter improved the recognition results for some speakers.

8. Results of research

During the author's research 500 records of digits from 0 to 9 spoken in Polish were used. They include men's, women's and children's voices. For all these records the envelope patterns were found. As the results show it was possible to find the envelope pattern for each digit (10 patterns) which are common for all speakers. Then the new method of the

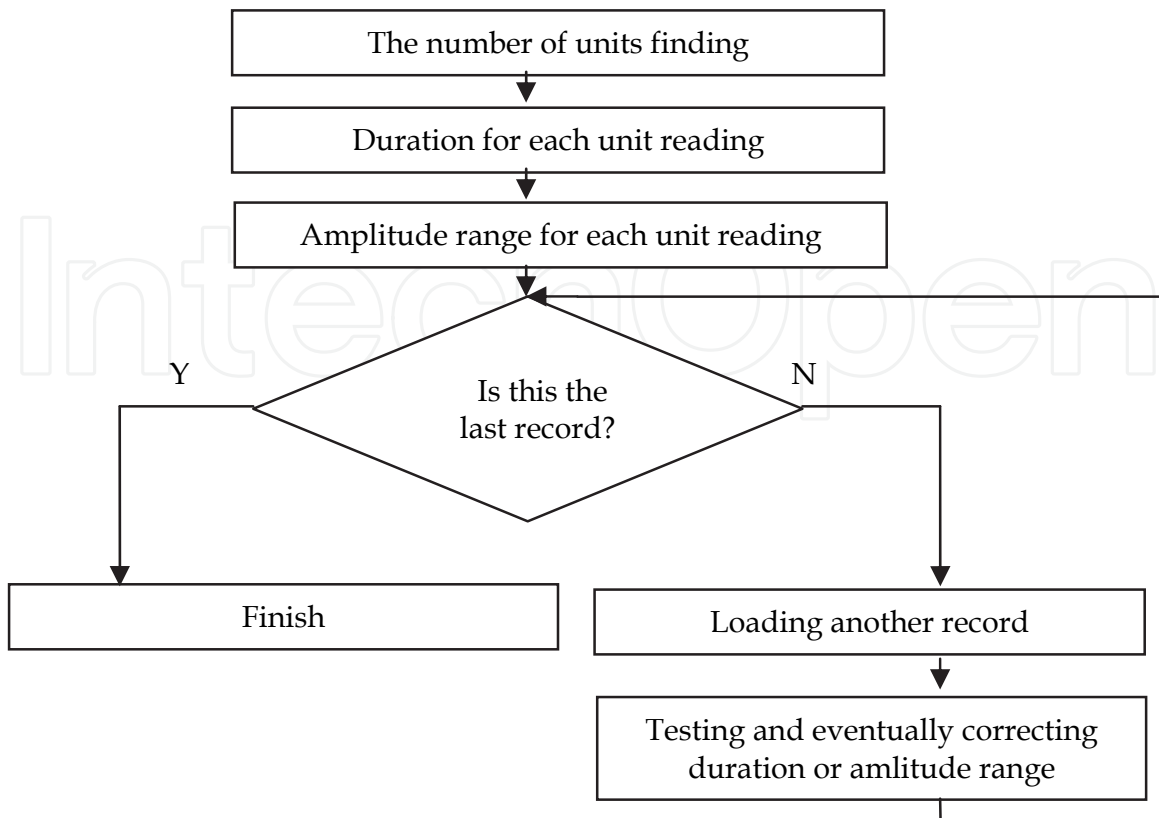


Fig. 17. The envelope patterns finding

phonemes recognition was worked out. It allowed to find voiced and noisy phonemes inside each word. From here the simplify phonemes structure for each word was obtained. At present this algorithm works properly for digits 0-9. In table 6 there are placed the results of recognition of author's system and HMM based system (Wydra, 2007).

Parameter	Author's method	HMM method
Number of recognizing words	10	20
Number of speakers	35	30
Recognition quality for digit "0"	100%	94%
Recognition quality for digit "1"	100%	98%
Recognition quality for digit "2"	100%	100%
Recognition quality for digit "3"	100%	100%
Recognition quality for digit "4"	100%	100%
Recognition quality for digit "5"	100%	98%
Recognition quality for digit "6"	100%	98%
Recognition quality for digit "7"	100%	96%
Recognition quality for digit "8"	100%	100%
Recognition quality for digit "9"	100%	100%

Table 6. Recognition results for author's and HMM based system for Polish.

As table 6 shows the recognition results of the author's system are the same or better than for HMM based system. Very important is also the fact that the recognition result in the author's system was achieved with lower amount of calculations and without a spectrum analyses. It means that it is faster and needs less memory and microprocessors operations, so it could be implemented in almost every simple microprocessor system.

9. Summary

Presented algorithm works properly for a small amount of word's systems. For bigger ones more parameters should be implemented. The new approach presented here is based on the electrical signal image recognition which is a source for all existing speech recognition methods. This signal includes all information necessary for reliable speech recognition. The main problem is what operation will be done in order to achieve the best recognition results. The most popular HMM method is based on the spectral and cepstral analyses which are complicated and difficult for implementation in simple processor systems. Here the signal processors should be used and conversion the signal from the time to frequency domain must be done. As the author's research showed, for small systems such as dialing a telephone number, changing the TV channel, choosing a level in the lift and many others, the systems based on the image recognition could be cheaper and more exact.

10. References

- Holmberg, M.; Gelbard, D.; Ramacher, U.; Hemmert, W.(2005). *Automatic speech recognition with neural spike trains*, Interspeech, pp.1253-1256, Lisbon, Portugal
- Hueber, T.; Chollet, G.; Denby, B.; Deryfus, G.; Stone, M.(2007). *Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips*, Interspeech, pp.658-661, Antwerp, Belgium
- Junho, P.; Hanseok, K.(2006). *A new state-dependent phonetic tied-mixture model with head-body-tail structured HMM for real time continuous phoneme recognition system*, Interspeech , pp. 1583-1586, Pittsburg, USA, 2006
- Ketabdard, H.; Vepa, J.; Bengio, S.; Bourlard, H. (2005). *Developing and enhancing posterior based speech recognition systems*, Interspeech, pp.1461-1464, Lisbon, Portugal
- Kumar, S.; Sreenivas T.(2005). *Speech Enhancement using Markov Model of Speech Segments*, Interspeech, pp.2069-2072, Lisbon, Portugal
- Nishida, M.; Horiuchi, Y., Ichikawa A. (2005). *Automatic speech recognition based on adaptation and clustering using temporal-difference learning*, Interspeech, pp.285-288, Lisbon, Portugal
- Seymour, R.; Stewart, D.; Ming, J.(2007). *Audio-visual integration for robust speech recognition using maximum weighted stream posteriors*, Interspeech, pp.654-657, Antwerp, Belgium
- Togneri, R.; Deng, L.(2007). *A structured speech model parametrized by recursive dynamics and Neural Networks*, Interspeech, pp.894-897, Antwerp, Belgium
- Vali, M.; Salehi, S.; Karimi, K.(2007). *Robust speech recognition by modifying clean and telephone feature vectors using bidirectional neural network*, Interspeech, pp.2554-2557, Pittsburgh, USA

Wydra, S. (2007). *Recognition Quality Improvement in Automatic Speech Recognition System for Polish*, Eurocon, pp.218-223, Warsaw, Poland, 2007

IntechOpen

IntechOpen



Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7

Hard cover, 432 pages

Publisher InTech

Published online 23, June, 2011

Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Janusz Dulas (2011). Speech Recognition Based on the Grid Method and Image Similarity, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, Available from:
<http://www.intechopen.com/books/speech-technologies/speech-recognition-based-on-the-grid-method-and-image-similarity>

INTECH
open science | open minds

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the [Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen