Janusz DULAS

Politechnika Opolska, Instytut Elektrowni i Systemów Pomiarowych

Automatic word's identification algorithm used for digits classification

Streszczenie. *Artykuł przedstawia efekt kilkuletnich prac autora nad stworzeniem algorytmu automatycznej identyfikacji cyfr wypowiadanych w języku polskim. Nowatorska metoda wykorzystująca analizę obrazów otrzymanych z charakterystyk czasowych wypowiedzi pozwala na osiągnięcie lepszych rezultatów niż stosowane powszechnie analizy widmowe. (**Algorytm automatycznej identyfikacji słów w wykorzystaniu do klasyfikacji cyfr**)*

Abstract*. This article describes the results of some years of research into automatic digits' identification algorithm for Polish. The new method based on the image recognition received from time characteristics gives better results than well known frequency domain analyses.*

Słowa kluczowe: Automatyczna segmentacja mowy, analiza czasowa, sterowanie za pomocą mowy, automatyczne rozpoznawanie mowy
Keywords: Automatic speech segmentation, time domain analysis, speech controlling, automatic speech recognition.

Introduction

Automatic speech recognition systems are more and more popular in our life. Now we have voice-controlled mobile phones, radios, information systems and many others. Nowadays the most popular method used for automatic speech recognition is The Hidden Markov Model (HMM) method [1,2,3] which is based on finding certain amount of states and the probabilities of transition among them. Another method which is not as popular as HMM method is the Neural Network application [4,5,6]. In this case The Neural Networks learn the right answers using many, mostly frequency domain parameters. The new method worked out by the author uses the images recognition received from the time characteristics of the spoken words. This is a new approach which doesn't use the frequency domain parameters but the results of recognition are the same or better than in other methods.

The grid method

This is the author's method where on each pitch period the rectangle grid is placed and the binary matrix is obtained. This process is shown in figure 1.

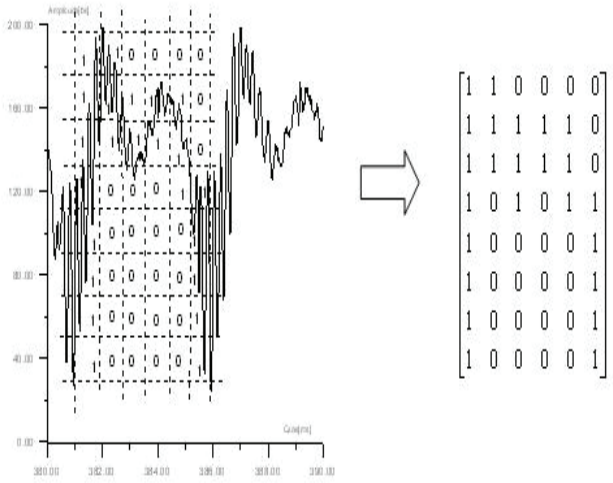

Fig.1. The grid placed on pitch period and binary matrix obtained from it.

The size of the grid is automatically fitted to the pitch period duration and the signal's amplitude. This way, the same word said with different amplitudes will give the same recognition result. Also fitting to pitch period duration is very important because each person has their own duration value. Usually women have a lower value than men. As author's research showed for the first group it is between 2 and 4 ms and for the second group it is between 4 and 11 ms respectively. Pitch periods are present in all vowels and some consonants so the grid method enables finding such phonemes. The next step in automatic recognition process is comparing each matrix with 5 previous and 5 next matrixes and calculating the similarity coefficient. If more than 88% of bits in compared grids are the same – they are treated as similar. The value of the similarity coefficient could be between 1 (the analysed grid is similar only to itself) and 11 (the analysed grid is similar to 5 previous and 5 next matrixes). Figure 2 shows the example of similarity coefficients obtained for the word "zero".

1,1,3,1,2,2,5,8,7,8,11,10,10,9,9,7,7,3,3,7,2,4,2,1,
2,3,2,1,2,1,1,1,1,1,3,2,3,5,1,2,5,4,3,4,3,5,4,6,5,6,
4,7,8,8,10,11,11,11,9,9,10,11,11,11,11,11,11,11,
10,9,11,11,11,11,11,11,11,7,11,11,10,9,8,2,4,4,1
,1,2,1,1,1,3,3,5,7,5,6,6,5,6,6,4,6,8,7,7,8,9,11,11,
11,11,11,11,6,7,7,7,7,7,8,8,9,7,8,7,6,7,2,4,3,3,2,

Fig.2. Similarity coefficients obtained for word „zero".

The groups of similarity coefficient's high value (usually equal or more than 3) mean the voiced phoneme and groups with lower values mean breaks between phonemes or transition zones. In figure 2 the voiced phonemes were matched with grey border. The first group represents phoneme "z" , the second "e" and the third "o". there is no "r" phoneme because there are no pitch periods there. More details about the grid method were published in [7].

The envelope analysis

For digits from 0 to 9 set the envelope patterns were found. They are common for all speakers. Each pattern consists of some amount of parts where each part has the defined minimum duration and the amplitude range. The examples of patterns for digit 0,5 and 9 are shown in figure 3.

The envelope analyses were used in automatic digits' recognition system. More details about the envelope patterns were published in [8].

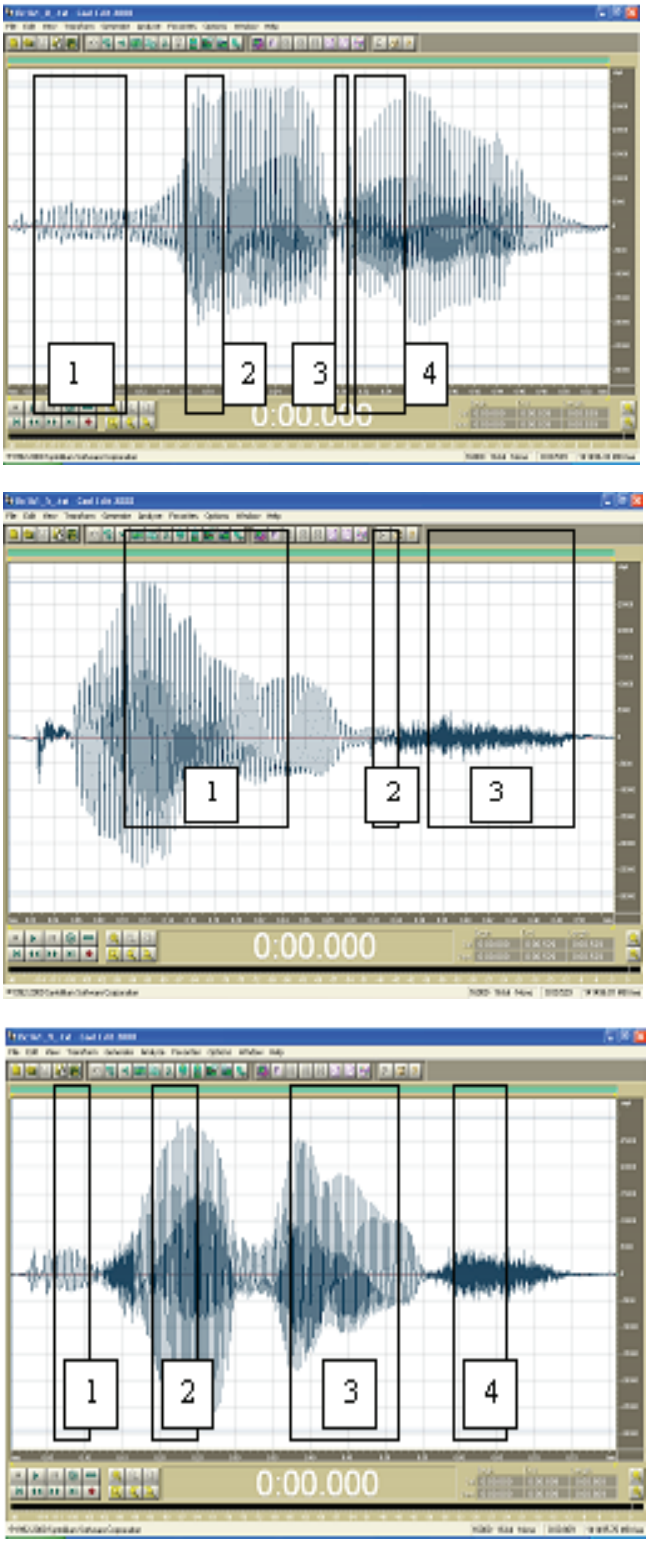

Fig.3. The envelope patterns for digits 0,5 and 9

The noisy phonemes' automatic localization

One of the phoneme group's are the noisy phonemes. They do not have any pitch periods and have big signal variety. In Polish digits they are included in 4,5,6,7,8 and 9. The examples of noisy phonemes in digits 4, 6 and 8 are shown in figure 4. As it is easy to observe, in one word there could be placed more than one noisy phoneme. As author's research shows, they could be found if in the speech signal for at least 100ms there are no pitch periods, the number of local minima is big (usually more than 200/100ms) and the signal amplitude is more than 5% of the maximum signal in the tested word. More details are presented in [9].

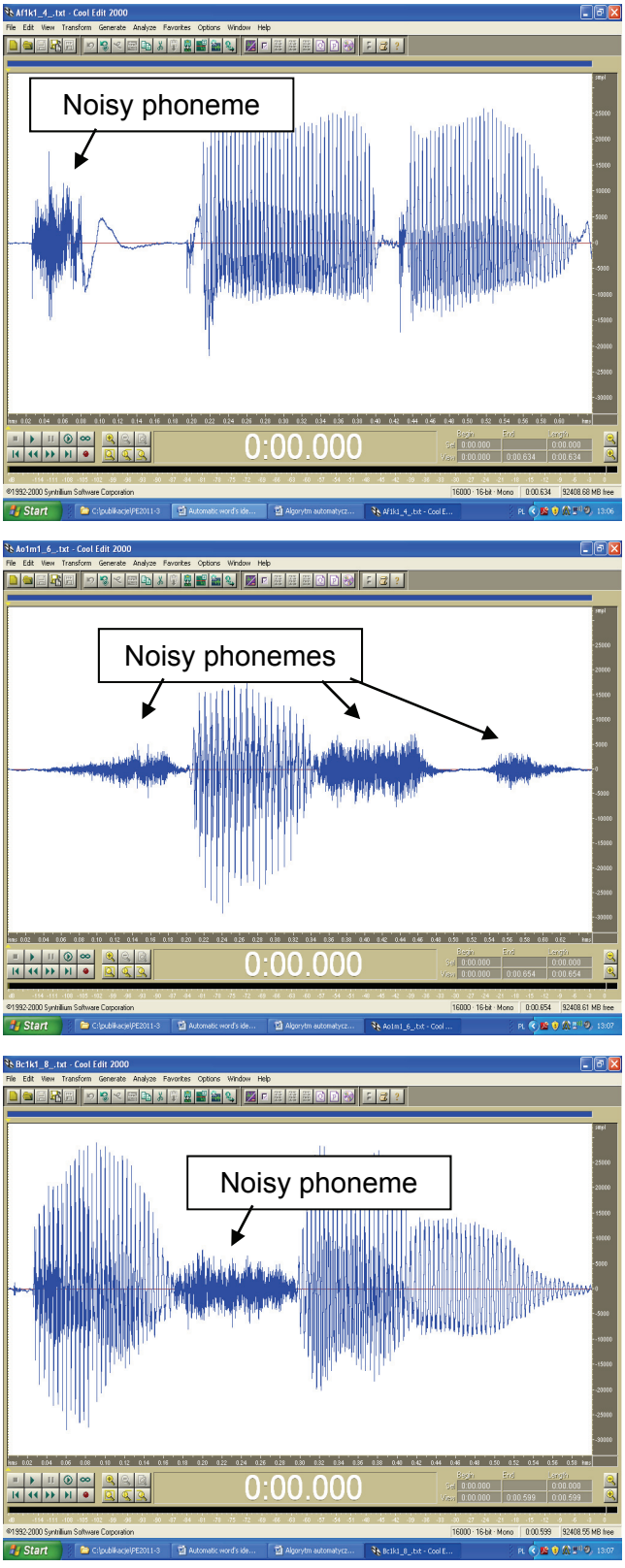

Fig.4. The noisy phonemes in Polish digits' names (for 4,6,8)

The algorithm of the automatic digits' names recognition for Polish

As the author's research shows, using voiced phoneme finding, envelope analysis and noisy phoneme automatic localization enable to build the automatic digits' names' recognition algorithm. It is shown in figure 5.

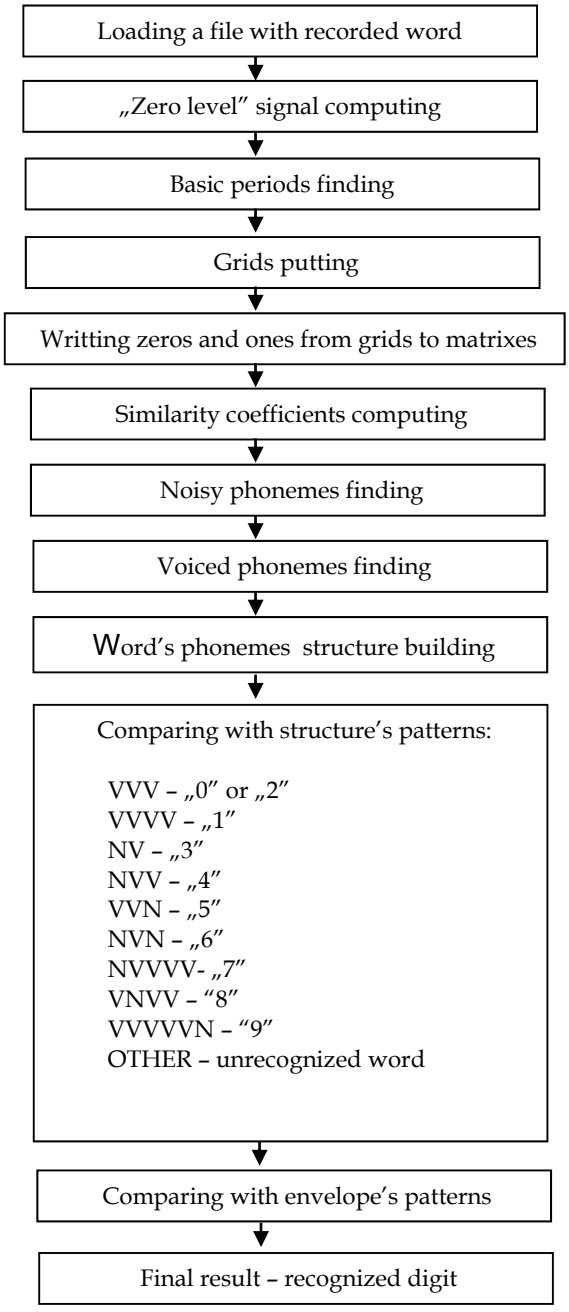

Fig.5. The simplified automatic digits' recognition algorithm

The first step is loading the file with recorded word. At present the *.txt file format is accepted but this is not a big problem to transform a file to other formats to *.txt.

Next step is a zero signal computing. Here 10ms signal before the recorded word is analysed and the mean value is computed.

The next operation is the basic period finding. Here the local signal's minima are observed. If there are no lower signal's values around the local minima, the signal's value is lower than the "zero level" and the time between such minimums is longer than 2 ms it means that the beginning of the next basic (pitch) period is found.

The next step is the grids' placing. If the beginnings of the following pitch period are found the grids placing is possible. On each basic period one grid is placed (Its size is fitted to the pitch period duration and to the signal amplitude).

The following step is coding the signal. Here in cells where there is signal the "1" is written and where there is no signal –"0". This coding result is written to the binary matrix.

The next operation is the similarity coefficient computing. Each matrix is compared with five previous and five following. This way the coefficient's value is received. Its value could be between 1 (there are no similar matrixes) to 11 (all matrixes are similar).

The noisy phonemes finding is the next algorithm's step. It is very important because it shows which matrixes could be used for voiced phonemes finding and which couldn't. Also the position and number of noisy phonemes is important for phoneme's structure building.

The next step is the voiced phonemes finding. Here the group of matrixes with high similarity coefficients are found. Also the groups with low similarity coefficients are analysed (they show the breaks between phonemes).

The next operation is word's phonemes structure building. Here voiced phonemes are matched by "V" and noisy phonemes by "N". Also their position in the word is important. This way the simplified phoneme's structure is obtained.

Comparing with structure's patterns is the next algorithm's step. For each digit's name spoken in Polish the simplified phoneme's structure is written. These patterns are compared with the analyzed word's structure. If any of them are the same it means that there is a high probability that the right recognition result was found.

The next operation is comparing with envelope's patterns. Here, ten patterns (for ten digits) are matched to the analysed word. The similarity is given in per cents. The highest result means the highest probability of finding the right digit.

The last step is finding the recognition result. If the structure of the word is the same as the one of the structure's pattern and the envelope shape is the same as envelope's pattern for the same digit, it means that the right recognition result was found. Otherwise, the algorithm gives the message "unrecognized word" or suggests the most probable right answer.

During the research, all records (500) were recognized correctly. As it was shown in figure 5, in block 'Comparing with structure's patterns' 9 digits' names out of 10 have different phoneme structures so only for differing digits "0" and "2" the envelope analysis is necessary. For others, it is the additional factor which tests the recognition's result correctness.

The block diagram shown in figure 5, shows only the main steps in automatic speech recognition algorithm. Each block consists of many operations but usually there are simple comparison operations which don't need much time to make. Using an image recognition method allows to avoid a time consuming frequency analyses and gives quite good recognition results comparing to well known methods used nowadays in many applications [10].

Summary

The new method of speech identification presented in this article is based on the image of the signal's time characteristic recognition. It was tested on 500 records said by people of different age and different sex. The results show that the quality of recognition is quite good and even better than the most popular methods used for frequency analysis. Although the research was made on the digits names from 0 to 9, the method could be useful for other words' recognition. In this case only more envelope patterns must be defined. In the near future, the on-line system of automatic speech recognition will be built.

REFERENCES

[1] Berkovitch M., Shallom D., HMM adaptation using statistical linear aproxximation for robust speech recognition , *Speech technologies, INTECH 2011, 303-320*

[2] KłosJuho P., Hanseok K.,A New state-dependent phonetic tied-mixture model with head-body-tail structured HMM for Real time continous phoneme recognition system , *INTERSPEECH 2006,* Pittsburgh, USA,1583-1586

[3] Weifeng L., Herve B., Non-linear spectral contrast stretching for in-car speech recognition, *INTERSPEECH 2007,* Antwerp, Belgium, 1122-1125

[4] Vali M., Salehi S., Karimi K.,Robust speech recognition by modifying clean and telephone feature vectors using bidirectional neural network, *INTERSPEECH 2006,* Pittsburgh, USA,2554-2557

[5] Holmberg M., Gelbard D., Ramacher U., Hemmert, Automatic speech recognition with neural spike trains, *INTERSPEECH 2005,* Lisbon, Portugal, 1253-1256

[6] Zouhour N., Laurent B. Frederic A., Comparison between two statio-temporal organization maps for speech recognition, Artificial Neural Networks in pattern recognition, *ANNPR 2006*, 11-20

[7] Dulas J., Speech recognition based on the grid method and image similarity, *Speech technologies, INTECH 2011, 321- 340*

[8] Dulas J., Automatyczna identyfikacja cyfr dla mówców polskojęzycznych, *PE 5/2010, 15-18*

[9] Dulas J., Szybka metoda identyfikacji fonemów szumowych występujących w cyfrach wypowiadanych w języku polskim, *PE 2/2011, 242-245*

[10] Wydra S. Recognition quality improvement In automatic speech recognition system for Polish, *EUROCON 2007,Warszawa, 218-223*

[11] Dulas J., Automatyczna segmentacja sygnałów mowy w oparciu o metodę siatek o zmiennych parametrach, *PE 1/2010, 229-232*

[12] Dulas J., Metoda siatek o zmiennych parametrach w zastosowaniu do rozpoznawania fonemów mowy polskiej, *Rozprawa Doktorska, Politechnika Opolska 2002, 36-42*

[13] Dulas J., Rozpoznawanie jednostek fonetycznych zawierających okresy podstawowe tonu krtaniowego, *Konferencja Podstawowe Problemy Metrologii, Sucha Beskidzka 2008*

[14] Dulas J., Analiza obwiedni jako parametr wspomagający automatyczną identyfikację wyrażeń, *PAK 5/2009, 308-309*

[15] Dulas J., Wspomaganie rozpoznawania wyrazów za pomocą opisu ich obwiedni, *Konferencja Podstawowe Problemy Metrologii, Sucha Beskidzka 2009, s.152-156*

[16] Dulas J., Automatyczne rozpoznawanie cyfr w języku polskim – identyfikacja fonemów szumowych, *PE 1/2011*

[17] Basztura Cz., Rozmawiać z komputerem, *Wydawnictwo Format, Wrocław 1992*

[18] Kłosowski P. Usprawnienie procesu rozpoznawania mowy w oparciu o fonetykę i fonologię języka polskiego, *Rozprawa Doktorska, Politechnika Śląska 2000*

[19] Nishida M., Horiuchi Y.,Ichikawa A.,Automatic speech recognition based on adaptation and clustering using temporal-difference learning, *INTERSPEECH 2005,* Lisbon, Portugal, 285-288

[20] Liu D., Kiecza D., Srivastava A., Kubala F., Online speaker adaptation and tracking for real-time speech recognition, *INTERSPEECH 2005,* Lisbon, Portugal, 281-284

[21] Xiang B., Nguyen L., Guo X. Fu D., The BBN Mandarin Broadcast News Transcription System, *INTERSPEECH 2005,* Lisbon, Portugal,1649-1652

[22] Lamel L., Adda G., Bilinski E., Gauvain J.L.,Transcribing lectures and seminars, *INTERSPEECH 2005,* Lisbon, Portugal,1657-1660

[23] Trancoso I., Nunes R., Neves L.,Recognition of classroom lectures in european Portuguese *INTERSPEECH 2006,* Pittsburgh, USA,281-284

[24] Chang-wen H., Lin-shan L., Extended powered cepstral normalization (P-CN) with range equalization for robust teatures in speech recognition, *INTERSPEECH 2007,* Antwerp, Belgium, 1106-1109

[25] Weifeng L., Herve B., Non-linear spectral contrast stretching for in-car speech recognition, *INTERSPEECH 2007,* Antwerp, Belgium, 1122-1125

[26] Seymour R., Stewart D., Ming J.Audio-visual integration for robust speech recognition using maximum weighted stream posteriors, *INTERSPEECH 2007,* Antwerp, Belgium, 654-657

[27] Zhu B., Hazen J., Glass R., Multimodal speech recognition with ultrasonic sensors, *INTERSPEECH 2007,* Antwerp, Belgium, 662-665

[28] Kacalak W., Majewski M., Inteligentny system obustronnej głosowej komunikacji system pomiarowego z operatorem dla technologii mobilnych, *PAK 4/2009, 221-224*

[29] Bekiarski A., Pleshkova-Bekiarska S., Pomiar sygnału głosowego za pomocą matrycy mikrofonowej dwuwymiarowej przeznaczonej do audio-wizyjnego sterowania robota, *PAK 10/2008, 741-743*

[30] Mięsikowska M., Narzędzie do przetwarzania i analizy sygnału mowy, *PAK 12/2007, 43-45*

[31] Mięsikowska M., Aplikacja umożliwiająca nawigację w Internecie za pomocą poleceń mowy, *PAK 5/2007, 87-89*

[33] Neiberg D., Ananthakrishnan G., Gołaś A.Blomberg M., On Acquiring Speech Production Knowledge from Articulatory Measurements for Phoneme Recognition, *INTERSPEECH 2009, Brighton, United Kingdom, 1387-1390*

***Autor**: dr inż. Janusz Dulas, Politechnika Opolska, Instytut Elektrowni i Systemów Pomiarowych, ul. Sosnkowskiego 31, 45-272 Opole, e-mail: j.dulas@po.opole.pl*