Dorota Kamińska
Instytut Mechatroniki i Systemów Informatycznych
Politechnika Łódzka

Artur Gmerek
Instytut Automatyki
Politechnika Łódzka

The open-source speech recognition system for robots control

The aim of the project was to develop an open-source speech recognition system, for controlling industrial and social robots. Particular deal of attention has been paid to the interoperability and speed of calculation in order to maintain real-time performance. Therefore, the source code is minimalistic, elastic and easily adaptable. Mel-frequency Cepstral Coefficents (MFCCs) combined with additional features and multistage semantic classifier have been applied in order to provide appropriate speed of calculation and classification accuracy. Dataset can be created by a user as a set of recorded words, which can be assigned to different classes (also specified by a user). Results of classification can be easily combined with specific robot commands. Those commands are sent to a robot through serial port, Telnet or TCP/IP protocol, accordingly, the system can be used with almost every modern robot. Our experiments have been conducted on Kawasaki FS003N robot. Source-code is available for download at: http://robotyka.p.lodz.pl/index.php?www=speechrob.html

1. Introduction

The goal of this project was to create a computer system (called *SpeechRob*) used for robots control, based on phrases recognized from human speech. Such system could be very helpful in everyday work with industrial and social robots. Moreover, it could advance control system to a higher semantic level. Communication with robots equipped with such a system is intuitive and natural, thus it can simplify any control operation, even for a first time user.

Machine control via speech could also be very helpful with emergency control. Robot could immediately react to some characteristic phrases, like "stop" and "hold", or similar spoken commands.

It was our goal to allow the system to cooperate with almost all types of robots, therefore Telnet and RS232 protocols (which are the most popular types of communication with robots) have been used. Basic parameters of both protocols are easy reconfigurable. Consequently, communication with most of robot types is possible and easily implementable. The designed system is characterized by a significant flexibility. Words that system has to recognize can be specified by a user and linked with selected robot action.

In order to maintain high accuracy and preserve speed of classification, all descriptors have been chosen with special care, mainly on the basis of scientific literature study.

Classification has been made using k-NN and pattern method algorithms, because of their high accuracy in relation to speed of calculation. One of our main goals was to enable natural human-machine interaction. Thus, a special search algorithm has been developed, to find characteristic words in an utterance. Additional information about this algorithm can be found in an appropriate section.

The code has been written in C# programming language, it is open source and available for download on the website *http://robotyka.p.lodz.pl/index.php?www=speechrob.html*. The authors hope that free license of program allows the researchers interested in signal processing to be able to use this code in their own projects and reconfigure it for their own needs.

2. Related Works

Speech recognition systems can be applied in numerous ways in science and in commercial products. One of the most advanced is *Dragon Naturally Speaking*, which is used for converting speech into text.

Several books have been written on this area of research. Some principal information about speech recognition can be found in [1]. In [2], statistical methods regarding speech recognition have been presented. Another good example can be found in [3], where speech recognition system in embedded systems and PC applications is presented. A superior example can be also found in [4], which depicts speech recognition issue from a practical point of view. Therefore some useful information is contained in [5], which connects speech recognition with discriminative learning. Moreover, in [6], some adaptive and hybrid methods have been presented.

There are however not many papers, related to connecting speech recognition systems with a robot controller. Lack of research in this field might be caused by high emphasis put on robot controlling solely by vision systems. In [7], authors have used particle filters for noise suppression to improve automatic speech recognition. Their experiments have been made using *Armar III* humanoid robot. The research has shown that usages of such filters have led to a significant increase of recognition accuracy. Another research on noisy environment cancellation for speech recognition enhancement human-robot interaction can be found in [8].

Special emphasis is also put on development of a real time robot controlling. This kind of systems should be optimized with respect to speed of execution. A good example of such work, which tries to deal with this problem, is [9]. System described in that paper is robust and able to work in real time. The article describes the problem of Voice Activity Detection (VAD). Another paper related with the above mentioned problem is [10]. It presents an automatic speech recognition system, which detects speech activity periods using Gaussian Mixture Model and MFCCs features.

Researchers generally use methods that rely on the analysis of individual phonemes. However, in this work the authors have classified whole words (divided into 3 parts). This method has been used deliberately, because such an analysis is faster when database consists of several dozen of words (as in this case).

Speech recognition for robot control is investigated by many scientists, yet still there are no free and open source applications dedicated to robot control based on speech recognition. For this reason the authors hope that this project will help other researchers in this field.

3. Methods

Studies have been carried out in accordance with the algorithm shown in Figure 1. First stage is connected with recognizing characteristic words. Phrases, situated near them, should be also recognized in order to send proper commands to a robot. Each of these words is searched for in different database. Statistical and spectral features have been used during classification. The classification has been done based on pattern method and k-NN algorithm. Main steps of the algorithm are described in following subsections.

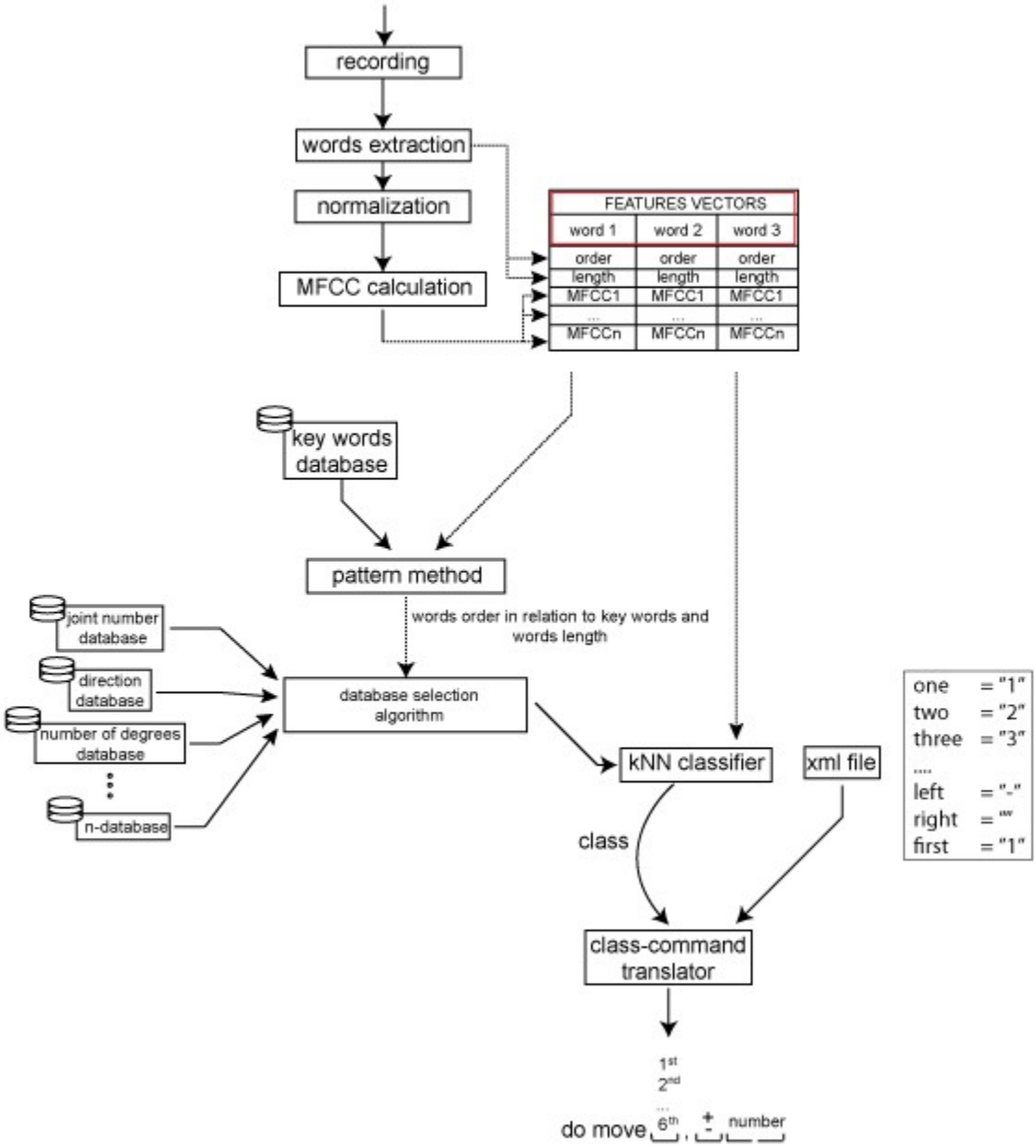

Figure 1: Sentence processing algorithm, responsible for spoken commands process

Database details

In developed system, training set can be created by a user. Nevertheless, for the purpose of these studies, built-in databases of English words have been prepared. Signals have been divided into four groups represented by different commands for the robot (key words, number of robot's joints, directions and number of degrees). Sample rate and format are not imposed and can be selected by a user. In this research all utterances have been saved in PCM WAV format file with 44100 Hz sampling rate. Samples have been recorded with the use of unidirectional microphone in an indoor environment. Experiments were performed on a laptop with dual-core microprocessor (Intel Centrino 2,6 GHz, 2GB RAM).

Words have been divided into classes and linked to different databases. One database with keywords corresponds to the other with detailed words. For example, while talking to a robot, one can use a phrase: "Robot, rotate by 10 degrees on the 3rd joint". Recognition of characteristic word "robot" activates the processing of the whole sentence. Firstly key words are found and their order is saved in a feature database. After that, analysis starts from word "*rotate*" and the system tries to find word which determines the angle of rotation, which is situated near the keyword - "rotate".

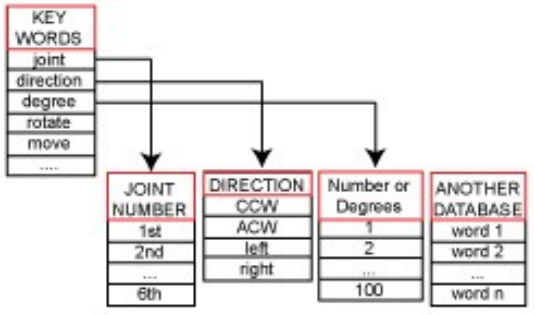

Figure 2: Relation between databases. Keywords from the first dataset correspond to other databases

Single words extraction

The main assumption of the described system is natural communication with a robot using complete sentences. However, the classification is based on single words extracted from spoken commands. Words extraction was performed by thresholding procedure of approximated original signal, which detects silence between each word in the process. The threshold is determined by the mean value of the speech signal (Figure 3).

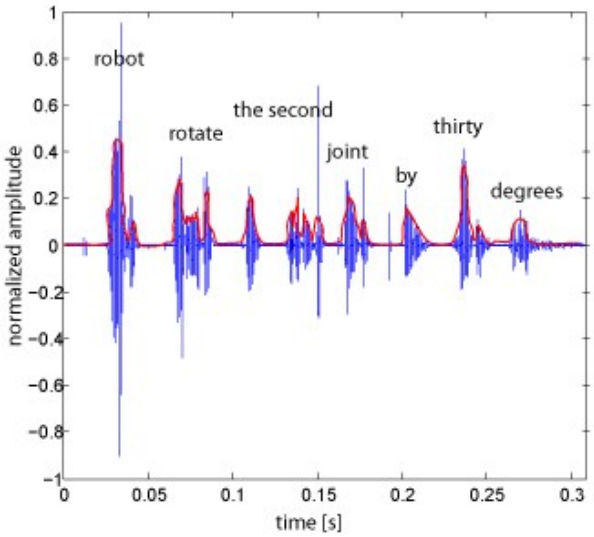

Figure 3: An example of a robot control sentence and approximation of the signal corresponding to it (red line)

Features extraction

Representation of the signal in time or frequency domain is complex. Therefore, features are sought to determine signal properties. In this part extracted features will be presented.

The usage of nonlinear frequency processing in automatic speech recognition systems leads to increase of their effectiveness. Therefore, Mel-frequency Cepstral Coefficients are currently a standard in speech recognition.

In this method, at first, the signal is multiplied by the Hamming window, presented by Eq. 1. In this research the duration of window was chosen experimentally and is set on 0.023 s.

$$w(n) = 0.53836 - 0.4616\cos\left(\frac{2\pi n}{N-1}\right) \tag{1}$$

where N specifies window size.

Subsequently, the Fast Fourier Transform is computed on each frame. Then the estimation of power spectral density function is calculated and averaged using overlapping triangular weight functions. Design of the triangular functions includes mel scale. The following equations present the frequency, m is frequency value in mel, f in hertz using natural logarithm, presented by Eq. 2 and 3.

$$m = 1127.01048\ln\left(1 + \frac{f}{700}\right) \tag{2}$$

$$\mathrm{f} = 700\left(e^{\frac{m}{1127.01048}} - 1\right) \tag{3}$$

The last step is calculation of a *Discrete Cosine Transform* (DCT) of the logarithmic estimation, using Eq. 4.

$$C_k = \sqrt{\frac{2}{L}}\sum_{l=0}^{L-1} ln\tilde{S}(l)\cos\left(\frac{\pi k}{L}\left(l + {}^{1}/_{2}\right)\right) \tag{4}$$

for k = 0,1, … q, where L is the number of weight functions, q is the number of Mel Coefficients.

In order to adjust automatic speech recognition (ASR) for k-NN and pattern method classifiers, the following method of features calculation has been developed. Each word has been divided on 3 equal parts. From each part 12 LPC coefficients have been calculated using a 23ms fragments. Results have been stored in the matrix. From each of such 3 matrices statistical parameters such as mean, standard deviation, maximum and minimum values were computed, giving 48 features as result from each fragment and 144 features, which characterized a single word. Subsequently, they were submitted to the process of classification using words from above mentioned databases as training set.

Classification

Classification is a statistical algorithm, which assigns objects to groups called classes, based on object features. The feature values, which are a source of information about the object, are usually presented by a vector:

$$x_j = [x^1, x^2, x^3, \ldots x^d] \tag{5}$$

where d is the number of features, is feature value.

All features values in a task are called the training set CU. The goal of classification is to assign a class i∈M for an individual object .

For classification in this system, mainly the k-NN algorithm was used, because of its simplicity and efficiency. In this algorithm the recognition process involves calculating distances in parameters space *X* between the unknown x_j object and all objects from the training set:

$$x_k \in CU \; for \; k = 1,2,\dots,I \tag{6}$$

where *I* is the number of training examples.

Various metrics $d(x_j,x_k)$ are used to calculate the distance. In this studies the Manhattan distance, presented by Eq.7 has been selected.

$$d(x_j,x_k) = \sum_{i=1}^{n} |x_{ij} - x_{ki}| \tag{7}$$

Obtained distances are sorted in the ascending order. Object is assigned to this class, which is the most common among *k* nearest objects.

4. Graphical User Interface

SpeechRob is an open source application to communicate with a robot using natural interface (speech). Moreover, it is equipped with a graphical interface, which allows user to adapt the system to the current needs. The GUI is divided into several parts, which are described below.

Samples recording interface

Recording tab (Figure 4) is connected with the first application part responsible for recording speech samples. It allows the user to collect an individual set of samples, from which a training set can be created. To record an utterance one must press the Start button and begin to speak to the microphone. A single word or a sequence of words can be recorded, the latter will be split into single words by the system. During recording of sequences, one should make pauses between each word. The Stop button ends the recording process. Furthermore, user is allowed to play the recording and plot a signal on a chart in the time domain.

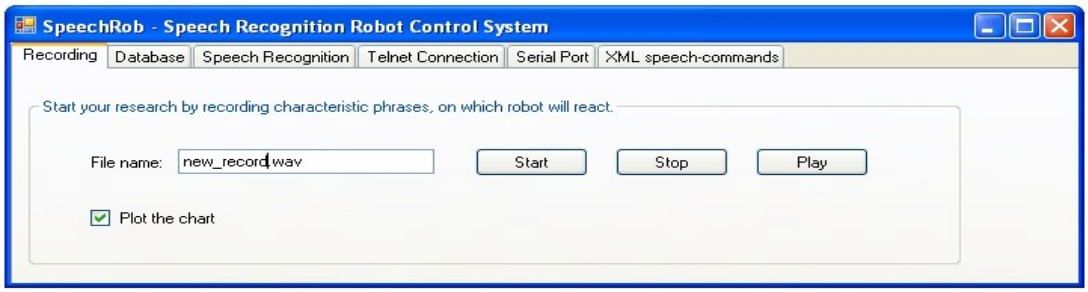

Figure 4: Recording interface used for sound files recording

Database interface

The second part of the application is responsible for creation of the training set. It can be found in Database tab (Figure 5). All previously recorded samples are visible in files list. One can freely add to, or delete recordings from particular class of the training set. Classes, which represent concrete words,

are also created in the described tab. After creating the training set, one can save it to a selected file using *Save and Extract* button. In the same time, features described in Section 2 are extracted from selected recordings and finally saved to .txt file.

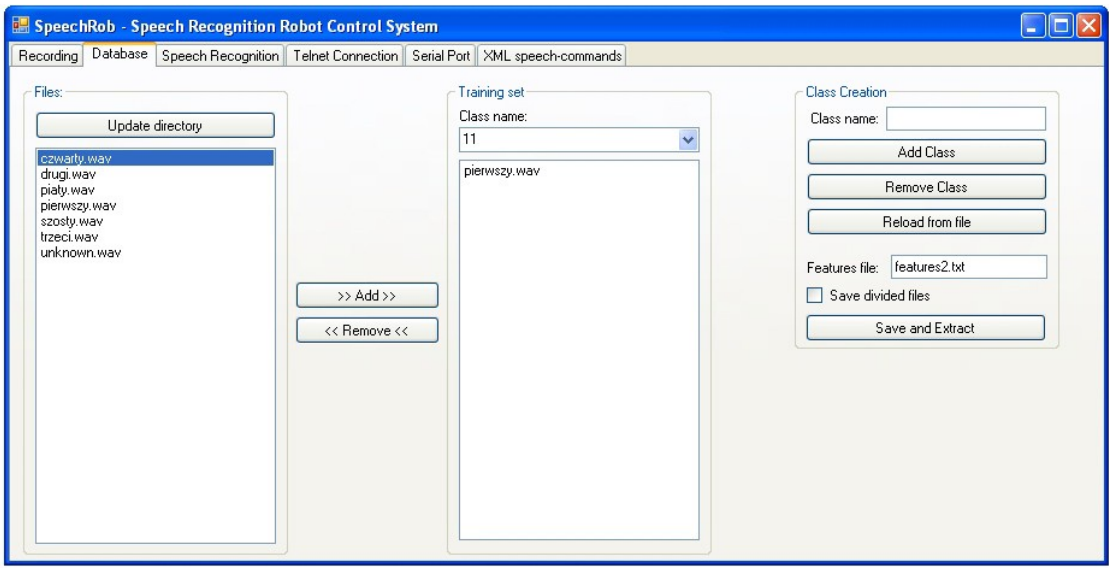

Figure 5: Database interface used for databases and classes creation

Speech recognition interface

In order to verify the quality of training set, speech recognition functionality was implemented (Figure 6).

Figure 6: Speech recognition interface used for spoken commands recognition process

One can record new samples or sequences, which will be classified by the system. The list of recognized words will be presented in the window. When the training set is efficient enough, application can be tested in real terms, in conjunction with a robot. The target functionality of this application named work in real-time has almost identical functionality, with a difference that the signal is being recorded continuously in real time and after processing the results are mapped to robot commands and further sent to robot controller. Before starting a recording user should have connection with a robot.

Communication interface

Next two tabs are responsible for communication with a robot. One can choose between Telnet (tab Telnet Connection), or RS232 (tab Serial Port) connection. In the first tab user can specify the IP address and port number, under which the robot controller is expected to be found and activate/deactivate the connection. If such an option is mandatory, one can enter username and password required for user authentication.

The second one enables the connection using serial port (Figure 7). To activate this connection one should specify the port number, baud rate, parity, stop bits and number of data bits. Data can be sent in one text or byte format. Moreover, basic commands can be received and print in text box via built-in console.

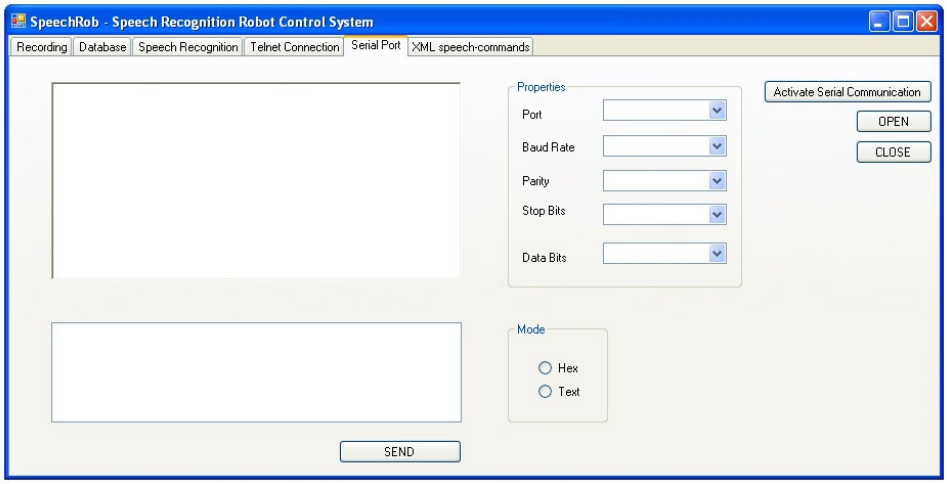

Figure 7: Serial port interface used for connection with robot via serial port

5. Experiments

Experiments have been performed on Kawasaki FS003N robot, which can be controlled with the use of AS language. *SpeechRob* system has been launched on the notebook, which has been connected to the robot controller by Telnet protocol. Microphone has also been plug in to the laptop.

In order to change the position of the robot, following command can be sent: *do drive 1,30*. The effect of this command should be the rotation of the first joint by 30 degrees in a clockwise direction.

Before controlling robot by the means of speech recognition, a class has to be created for each world, on which robot should react. For each of those classes a set of 40 words has been recorded by two different speakers. All classes have been divided into 4 databases. After training set recording, automatic recognition algorithm can be started. System records signal in a loop, defined by timer

period (2,5 sec.). Signal is then processed in order to check if user said "*robot*", "*stop*" or "*hold*". Word extraction algorithm and feature calculation have been used, which are described in methods section. These words are classified with the use of pattern methods algorithm. Patterns represent middle point of each class, therefore there are average features vectors calculated as arithmetic mean of the features vectors of certain classes. Classification is done on the open set, accordingly it is essential to check accuracy of classification defined based on Euclidean distance from the center of clusters. If the distance is lower than established threshold, the system classifies new word to one of the classes. When word "*stop*" or "*hold*" is detected, immediate command is send to robot in order to hold it. If user says "*robot*" and system detects it, the whole process of recording and processing of the spoken commands will be started (Figure 8). This principal algorithm of recognizing these three characteristic words (*robot*, *hold* and *stop*) is fast and lasts about 300ms (Windows 7 environment).

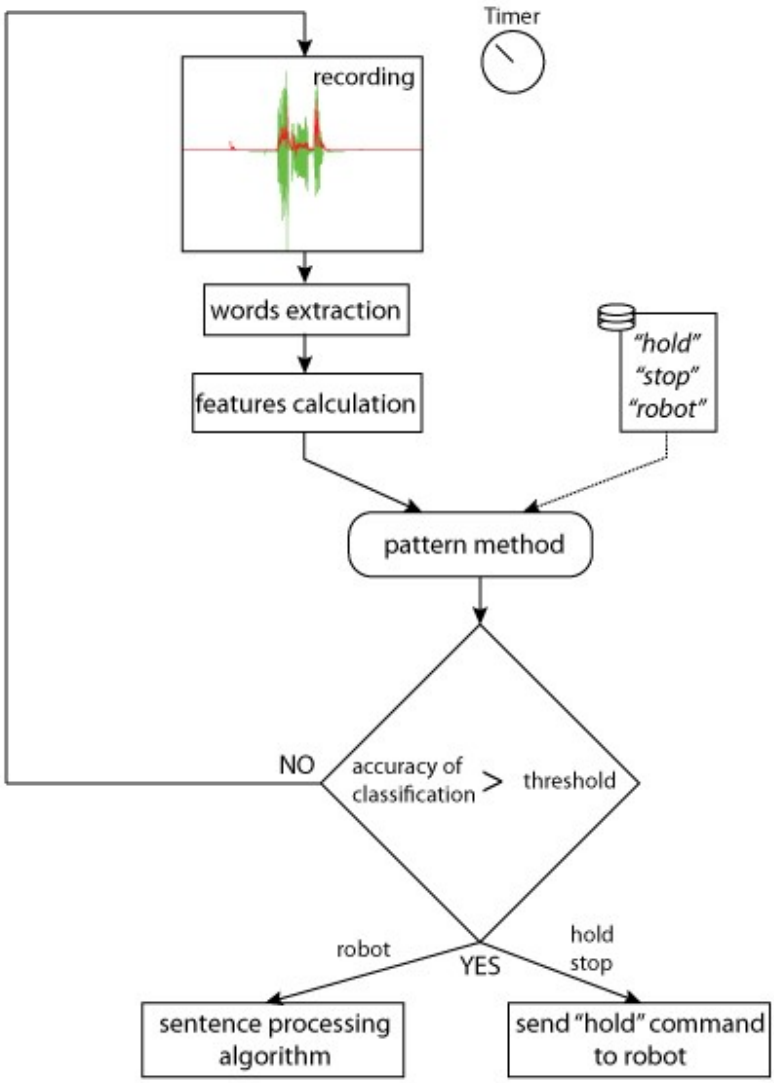

Figure 8: The main control algorithm. Algorithm is responsible for immediate holding a robot and starting to record spoken commands

After detecting the word "*robot*", command recording will be started. Recording stops after period of 10 seconds. Obtained signal is processed accordingly to the algorithm described in the methods section, which leads to recognition of spoken commands, which are connected to corresponding robot actions. After that process, the whole command is sent to a robot, what results in motion.

Commands corresponding to classification of words are written in an xml file, in which specific words are connected to robot commands. A connection between the word “one” and it’s numerical equivalent can serve as an example of such a connection. This file can be easily changed in order to provide another type of robots control.

Figure 9: Kawasaki FS003N robot, on which the experiments have been carried out

6. Discussions and conclusions

This paper presents system used for robot control, based on speech recognition. It is adaptable and able to control different robot models. The program can be connected to robots based on Telnet or RS232 standards. A user can easily create database by recording characteristic phrases before automatic work. A great deal of effort has been put into the process of creating a system which is able to quickly process natural-speech sentences spoken by a user.

Used classification method is unusual in case of speech processing. Scientists usually used HMM classifiers. However, creating a HMM model of words from the phonemes is a time consuming process, because one must not only create a model, but also perform words to syllables division prior to model creation. In the case of a database containing a low number of words, the k-NN classifier with a vector of described MFCC features is a more efficient method than a HMM classifier when considering memory usage and time of calculation (accuracy of classification of both is similar). Accordingly it could be used for example in handheld devices.

Research has shown how to create dynamic, reconfigurable and fast speech recognition system for robots control. Mean results of the k-NN classification done on close data set have reached 97%. Average speed of the system, from the uttered sentence to the performing of commands by robot, was one second.

The idea of connecting key words to different databases is also worth mentioning. The concept comes from the fact that certain words will always occur close, in a sentence, to other related ones. For example the words for direction (“left”, “right”) or rotation (“clockwise”, “counterclockwise”) will

always be close in the sentence to words “direction” and “rotation”. This method greatly accelerates databases searching process and command recognition.

7. Future works

In the future, the database selection algorithm will be expanded. Also multithreading will be implemented, which will greatly reduce system reaction time. Authors would also like to implement robot’s reaction to undefined or missing phrases. If some words, which are not defined in training set, appeared in the sentence, the user would be asked to explain the unknown phrases. Moreover, if the user omitted any keywords necessary to determine a concrete control sequence, he would also be asked for additional information.

Finally the system will be included into a software robotic platform, in order to communicate with other robot subprograms. Using robotic platform is a contemporary trend, which simplified communication in social robots between modules [1].

8. Acknowledgments

Artur Gmerek is a scholarship holder of grant financially supported by the Ministry of Science and Higher Education of Poland (Grant No. N N514 469339).

Authors are grateful for valuable insight into this work and advice to prof. Adam Pelikant and prof. Edward Jezierski from Technical University of Lodz.

References

[1] R. Tadeusiewicz. *Speech in human system interaction*. In Proc. 3rd Conf. Human System Interactions (HSI), pages 2–13, 2010.

[2] Frederick Jelinek. *Statistical Methods For Speech Recognition*, ISBN 0-262-10066-5, Massachusetts Institute of Technology 1997.

[3] Jean-Claude Junqua. *Robust speech recognition in embedded systems and PC applications*, ISBN-10 0792378733, Kluwer Academic 2000.

[4] LucioPrina Ricotti Claudio Becchetti. *Speech recognition: theory and C++ implementation*, ISBN-10 0471977306, Wiley 1999.

[5] Xiaodong He Li Deng. *Discriminative Learning for Speech Recognition, Theory and Practice*. Morgan & Claypool, 2008.

[6] Herve A.Bourlard Nelson Morgan. *Connectionist Speech Recognition, A Hybrid Approach*. Kluwer Academic, 1994.

[7] F. Kraft and M. Wolfel. *Humanoid robot noise suppression by particle filters for improved automatic speech recognition accuracy*. In Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems IROS 2007, pages 1737–1742, 2007.

[8] Sheng-Chieh Lee, Bo-Wei Chen, and Jhing-Fa Wang. *Noisy environment-aware speech enhancement for speech recognition in human-robot interaction application*. In Proc. IEEE Int Systems Man and Cybernetics (SMC) Conf, pages 3938–3941, 2010.

[9] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J.-M. Valin, K. Komatani, T. Ogata, and H. G. Okuno. *Real-time robot audition system that recognizes simultaneous speech in the real world*. In Proc. IEEE/RSJ Int Intelligent Robots and Systems Conf, pages 5333–5338, 2006.

[10] C. T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, and N. Hagita. *Robust speech recognition system for communication robots in real environments*. In Proc. 6th IEEE-RAS Int Humanoid Robots Conf, pages 340–345, 2006.

[11] Artur Gmerek. *High-level controller for an arm rehabilitation robot - positioning algorithms with respect to EMG data*. In MMAR, pages 182–187. IEEE Proceedings, 2011.