Control of Assistive Tools Using Voice Interface and Fuzzy Methods

Vytautas Rudzionis, Rytis Maskeliunas, and Tomas Rasymas

Department of Informatics, Vilnius University Kaunas Faculty, Lithuania
Kaunas University of Technology, Lithuania
vytautas.rudzionis@vukhf.lt, rytis.maskeliunas@ktu.lt

Abstract. The paper describes voice controlled multimodal assistive system with fuzzy control. Voice commands are often most convenient way to control various assistive tools. For full functionality voice commands need interpretation. The detection of voice boundaries in the long audio recording was implemented. The experimental results of fuzzy based indoor navigation system are presented in this article. The fuzzy control strategy presented bellow works on a given trajectory principle. The position of the device, the distance from the trajectory, orientation and control tasks are evaluated according to visual data. Paper presents the control model and algorithm of a real-life prototype.

Keywords: voice technology, voice command recognition, voice user interface, fuzzy methods, intelligent control.

1 Introduction

The speech is often the most natural, easiest to use and the most convenient way of human-machine interaction. Implicit richness of human speech communication gives the user many degrees of freedom for control and input of various devices. Voice user interfaces has a series of advantages for the control of various applications. The advantages of voice user interfaces were described and shown in various papers such as [1].

In recent months personal digital assistant Siri introduced by Apple as an integral part of iPhone4 series attracted a lot of attention. The essential property of Siri is user interface controlled by voice commands. The success of Siri is caused mainly by three components: improved speech recognizer, improved natural language modeler and semantic analyzer. All these three components enabled the implementation of voice user interface with the flexibility never seen before in practical applications available to the general public. It should be emphasized that speech recognition accuracy is the first and basic element of such types of systems since only good enough voice recognition will enable to implement semantic analyzer. Another very important characteristic of Siri is placed even in the name – personal digital assistant. It could be speculated that voice controlled user interfaces could be applied as a basis of personal assistive tools for the people with disabilities too.

W. Abramowicz et al. (Eds.): BIS 2012, LNBIP 117, pp. 308–318, 2012.

Applications of speech technology can be grouped in the areas of access, control, communication and rehabilitation\therapy. For people with different impairments different types of speech technologies are more important: for people with visual impairments speech synthesis is essential as a way to access information, for people with hearing impairments perceptual speech processing and amplification are crucial, for other disabilities other areas of speech technology can be more important. But it is really difficult to find people with some sort of impairment that cannot benefit from one or another aspect of voice technology.

The main group of interest which needs is addressed in this study is the motor-handicapped people. The characteristic property of such category of people is that they often simply can't use traditional keypad based control systems independently or the use of such systems is significantly restricted. Environmental Control Systems (ECS) or Smart home control interfaces are available which address many elements of home management for disabled people, such as control of audio-visual equipment, telephones, household appliances, doors and curtains as well as the ability to summon assistance. Most ECSs utilize switch-scanning or keypad interfaces for control. More recently, ECSs with speech recognition have been introduced and a number of such systems are available on the market. Their success depends on a number o factors most important of them being maturity of voice processing technology used. Even better results could be achieved implementing multimodal approach – combining several different modalities to work in parallel or supplementing each other. In example, a multi-modal interaction framework using speech recognition and computer vision to model a new generation of interfaces in the residential environment was developed in [2]. The design is based on the use of simple visual clues and speech interaction. The latter system incorporates video information processing block which moves this system to the class of multimodal systems. Experience shows that motor-handicapped people are keen to use voice technology. This is especially true for people with hand movement restraints were the use of voice recognition is the only mode to transfer a computer control commands.

This paper presents a study about the possibilities to implement intelligent control methods for the one of the most important assistive tools used by the disabled persons – the wheelchair. The wheelchair is controlled using voice commands. The very important factor is the quality of the recognition. Paper deals with the accurate detection of the speech utterance boundaries during continuous audio input. Some of the voice commands need to be interpreted. E.g., such command as "turn to the right" requires find the necessary angle for the rotation. The angle depends on the location of various objects in the proximity of wheelchair. In such situations fuzzy based intelligent control methods are used to find the necessary decision in the scene. Chapter 2 of this paper describes the investigation of detection of speech boundaries in long audio recordings. Chapter 3 deals with the issues related with the implementation of fuzzy methods for wheelchair movements control.

2 Recognition of Voice Commands Using the Detection of Acoustic Events in the Long Utterances

One of the characteristic properties of voice user interfaces targeted to the disabled people controlling such assistive tools as the wheelchair or similar device is the necessity to operate in the continuous audio input mode. Most of the commercial tools implementing voice control (e.g. SIRI personal assistant manager) are using manually controlled audio interface: user needs to press button and then to speak (to say voice command) and often to release the button to indicate that utterance has ended and recognition should be started. In principle this approach may be implemented for the disabled people too but often it is unacceptable. People with some kind of disabilities such as motoric disabilities can't move hands easily or may have difficulties when pressing the necessary button or key. In this case we need to implement continuous audio mode: audio signal is recorded permanently and passed to the signal processing system for further analysis continuously.

Here we have two choices: first choice is to pass to the recognition system each part of the recorded audio signal and try to recognize acoustic-phonetic content of the recording (even to recognize the noise and some nonsense recordings) or to try to find those parts of audio recording were user of the assistive tool is speaking and to pass only those parts of the recording to the speech recognizer. The first approach is more likely to produce more recognition errors since the bigger variety of audio patterns needs to be recognized while the second one has the potential to be more accurate. But for the second approach to be accurate enough important issue to detect properly the places were the user of the system is speaking should be solved.

The detection of the boundaries of acoustic events such as utterances in the long recordings, utterances in the noisy environment or the phoneme boundaries within a word is one of the most fundamental problems in speech processing. A lot of activities were devoted to solve this problem. Various algorithms were proposed for the detection of speech and segmentation of spoken utterances, e.g. several methods could be found in [3-6]. Most of the algorithms exploit such spoken signal properties as the articulatory movement features or the differences between the actual signal spectrum and the spectrum prediction using its first or second order regression. Many methods also exploit signal energy changes as the factor. The selection of those features are based on the analysis of the physical properties of speech signal, e.g. articulatory movements features describe the particular structure of the speech signal spectrum which is typical only for the transitions between different acoustic events.

We proposed proprietary speech detection in noisy and long recordings method [7]. This algorithm proved to be accurate and robust enough for the segmentation of spoken speech for a wide range of SNRs and various classes of noise types. In [8] was used modified algorithm for the task of the acoustic event segmentation. The algorithm described here is further modification of the algorithm presented in previous study.

The speech is non-stationary process over longer time spans. The non-stationarity of speech signal is the result of different nature of different phonetic units. At the same time speech could be considered as a quasi-stationary process over shorter time periods (a time frame is shorter than 30 ms though it depends on the phonetic content of a signal). The basic idea of speech detection algorithms of is to find the places in the long recording were the statistical properties changes rapidly enough.

The algorithm for the detection of the acoustic events boundaries consists of several steps:

1. The logarithmic spectrum obtained from Fourier transform, linear prediction coding and IIR filter bank is derived using 8-10 ms step and 25 ms analysis window. If we will compare this algorithm with the algorithm used in earlier study we should note that the combination of three types of different spectrum was used. Logarithmic spectrum vectors were used to construct the likelihood function of the spectral rate changes in the utterance. The changes in the likelihood function values are used as the main indicator of speech presence at a particular time moment.
2. Then the spectral change rate function was filtered and integrated over experimentally set time spans. The integration allows obtain smoother likelihood function and helps avoid the random type of fluctuations which are characteristic for the changes in spectral properties of many phonetic units. The sequence of filtered and integrated parameters was used to detect the boundaries of acoustic events and was called an acoustic events response (AER): if it value exceeds experimentally defined threshold we will fix the start of the utterance and if the value it will drop below the threshold level for a experimentally set time period we will fix the end of utterance.

The Figure 1, which from top to down consists of the oscilogram of the original and the differentiated syllable, the spectrogram of the same syllable and the AER curve, is used to illustrate the efficiency of algorithm. It is expected that the changes in the acoustical content of a signal will occur on the places where the AER curve reaches the local maximum. The higher is the peak of the AER the higher is the likelihood value of the boundary between the different acoustic units.

We investigated the algorithm described above experimentally to find its efficiency in the voice commands recognition task. One group of experiments was carried on with a set of commands applicable for the control of wheelchair and using commercial speech recognition tools adapted to recognize Lithuanian voice command. Another group of experiments was carried on trying to define the limits of method. In this case a phonetically complicated material was used.

In the first group of experiments utterances of ten digit names (0-9) and a set of 24 voice commands that may be applied to control assistive device were used. The digit names were pronounced by 20 different speakers and each command was pronounced 20 times by each speaker. 24 control commands were read by 16 different speakers. Each speaker read every voice command 20 times as in the above case. Also the possibilities to adapt Microsoft commercial English and Spanish recognition engines were investigated in this context.

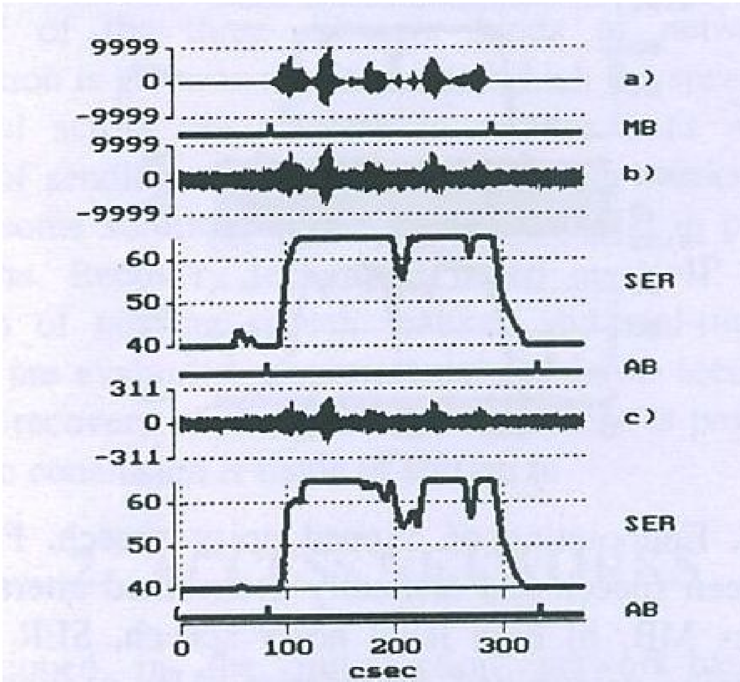

Fig. 1. Speech detection for various SNR. From top: clean speech and manually detected boundaries; low SNR signal, detected speech, higher SNR level, detected speech.

The voice command recognition experiments were carried on as follows. Long audio recordings were prepared (length of the recording is about 2 min). In each long recording voice command was inserted at a randomly selected place. The average length of voice commands was about one second and none of the tested commands was more than 2 sec in duration. To evaluate the impact of noise the additive white noise was added to the recording. In one case all recording was transferred to the recognizer and recognizer tried to recognize the acoustic content of whole recording. In the second case the whole recording was transferred to the utterance detection algorithm which tried to detect the part of the recording with the voice command present and only the detected voice command fragment was passed to the recognizer.

The Microsoft Speech API based commercial recognizers were used in this experiment. The English and Spanish recognition engines were used. For the adaptation to recognize Lithuanian voice commands methodology presented in [9] was used.

Table 1 shows the summarized results of this group of experiments. Here EN 7.0 means Microsoft English engine (ver 7.0) with ARPAbet-based phonetic transcriptions [10], ES 8.0 means that Microsoft Spanish engine (ver 8.0) with UPS-based transcriptions was used. Also Microsoft Speech Server English and Spanish engines (ver 9.0) (marked as EN9.0 and ES9.0 in the table) were used with UPS transcriptions.

Table 1. Recognition accuracy of voice commands in different environments and using different recognition engines (accuracy in percent)

Type of engine	Only voice command	Clean speech, SNR=30dB		Noisy speech, SNR=10 dB	
		Whole recording	Detected command	Whole recording	Detected command
EN7.0	87.4	79.3	85.6	58.6	78.8
ES8.0	94.6	88.3	92.3	66.5	85.6
EN9.0	77.0	74.5	76.8	51.2	63.4
ES9.0	97.0	75.6	97.0	80.4	92.3

The main conclusion which could be drawn from this table is that detection of voice commands helped to achieve higher overall recognition accuracy. The benefits of the voice command detection could be seen in all cases under investigation. Another observation is that the more complicated and noisy the acoustical environment is the more benefits could be get from the detection of voice command in the long utterance: if in the case with clean speech command detection allowed to achieve recognition accuracy increase by 10% in average then in the case with noisy speech increase in recognition accuracy was about 25% in average. Another observation is that better the base recognizer is (earlier studies showed that Spanish recognition engines are better suited for the adaptation to recognize Lithuanian voice commands than English engines; the reason is more similar to Lithuanian acoustic-phonetic structure of Spanish than English) the lesser the degradation of recognition accuracy in noisy environment could be expected.

Another group of experiments were performed using acoustically complicated and confusing data. These were utterances of syllables MA,NA,MI,NI,MO,NO known for their complicated phonetic structure and significant acoustic similarity to each other. Often such utterances are used to evaluate the potential of speech recognizer (so called boundary conditions). In these experiments phonetic material of 60 different speakers was used. Each speaker pronounced each syllable two times in isolation. One of those utterances was used for training (total 60 utterances training set) while another for testing (total 60 utterances testing set for each syllable). All utterances were manually processed marking the start and the end of acoustic phenomena. So the exact boundaries of each phonemic unit were known during training while for the testing they were known but weren't used.

In this case for the recognition was used CD-HMM based recognizer. The basic properties of HMM recognizer were straightforward: three states for each syllable and typical Baum-Welch reestimation procedure used to train the HMM model for each type of syllable. For the recognition also typical Viterbi search procedure was used. As for the acoustical front-end MFCC features were used to describe the acoustical structure of signal. MFCC features were supplemented with the change rate and acceleration coefficients (delta and delta-delta features) together with energy and energy delta. So the 39 features vector has been used. Output probabilities were modeled with single Gaussian and mixtures of several Gaussians.

Other methodology used in these experiments was the same as in the above described case: manually labeled clean recordings were used to get the reference recognition accuracy; later utterances of syllables were inserted into the relatively long recordings at the random places. These recordings were transferred to the utterance detection algorithm and detected part then used for the recognition. Also whole utterances were used as above. In this case HMM model chains silence-syllable-silence or noise-syllable-noise was used. Table 2 summarizes the results obtained during this group of experiments.

Table 2. Recognition accuracy of nasalized syllables in different environments and using different recognition engines (accuracy in percent)

HMM model type	Manually labeled syllable	Clean speech, SNR=30dB		Noisy speech, SNR=10 dB	
		Whole recording	Detected command	Whole recording	Detected command
Single Gaussian	52.4	44.7	48.8	33.6	39.5
2 Gaussian mixtures	67.8	48.7	61.0	41.1	48.7
3 Gaussian mixtures	79.7	62.1	75.6	49.6	55.6
4 Gaussian mixtures	85.6	67.7	81.2	67.5	74.3
8 Gaussian mixtures	89.4	78.4	86.5	71.2	80.5

As could be seen from the results in Table 2 implementation of voice detection algorithm had significant impact to the recognition accuracy. Recognition gains were obtained for each syllable under consideration and both in the clean and noisy environments. The bigger gains were achieved in the presence of stronger noise and when simpler recognition algorithm was used (e.g. fewer Gaussian distributions were applied for the output probability modeling).

3 Fuzzy Methods for the Control of Assistive Tools

Control of assistive devices using voice commands is attractive solution. But only proper recognition of voice commands can't solve the needs of disabled people immediately in many cases. Let us imagine the control of movement of wheelchair by voice: the person in the wheelchair wants to make turn to the left. The key issue is how to select the desirable angle for the turn (it is obvious that the command "turn to the left" will not mean always turn by 90 degrees). It is possible to introduce several commands for making left turns or even to introduce the possibility to say exact angle in degrees by voice but all of these solutions aren't convenient in many cases: often it is difficult to determine the exact angle for the turn immediately and it is particularly inconvenient to correct inaccurate decisions later.

In our multimodal assistive device control platform we tried to introduce fuzzy methods using intelligent controller to find the appropriate (more exact) decision in particular environment. This decision depends on the environment and the position of other objects in the nearby space. It means that when a command (let say "turn to the left") has been recognized then simplified analysis of surrounding space is performed using video streams from cameras mounted on the head of the user and the distances to the objects are found. The intelligent controller based on fuzzy methods makes a decision what a turn angle should be. Below we will describe the fuzzy based intelligent control system used in multimodal platform to control wheelchair for disabled person.

It should be noted that a numerous attempts to implement intelligent methods for movement control of robotic devices were done in the past. Several examples of such attempts could be found in [11-14]. The fuzzy methods were selected because of their flexibility and good results achieved in various applications of similar kind and complicity as well as our experience working with them.

3.1 Control Model for the Assistive Platform

In this paragraph we present the kinematic model of the mobility platform illustrated in Figure 2.

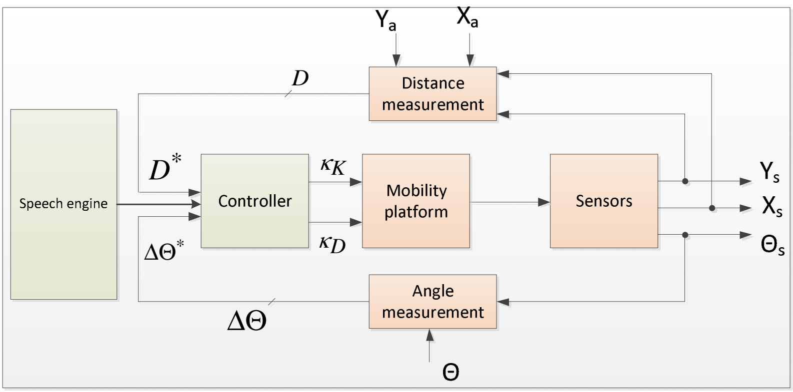

Fig. 2. Kinematic model of the wheelchair mobility platform

Think of it as a square with two sides capable of turning left and right by having a power vector on each side. The main movement rules mathematically could be written as follows.

If $P_D = s \times K_{P_D}$ and $P_K = s \times K_{P_K}$ then $K_P = \frac{(P_D - P_K)}{I}$ and $P = \frac{(P_D + P_K)}{2}$.

Here P – is a linear velocity, K – is the left side of the platform, D – is the right side of the platform, κp – is the angular velocity, I – the length of an axle, s – the radius of a wheel. The dynamics of this system are described using the state space variables. The movement of our mobility platform can be described like so $K_D = \frac{1}{S}P + \frac{1}{2S}K$ and $K_K = \frac{1}{S}P - \frac{1}{2S}K$. The control of which is further modeled by voice.

3.2 Fuzzy Control Rules

The main purpose of the rules described below is to find the angle of rotation in depending on the location of objects in the environment used to move by the wheelchair. Naturally the system we are describing is non-linear, because of various factors such as time delay, noise, channel effects and other factors. The orientation θ of the device is used as second input for the fuzzy controller and it shows how the device is oriented with respect to the trajectory segment. The orientation can get values from interval [-180^0, 180^0]. The mobility device is moving in good direction when orientation value is equal to 0. The device is moving towards trajectory when θ is has a positive value, otherwise it moves from trajectory when θ has a negative value. The angular velocity of the right power vector must be higher than the angular velocity of the left power vector, when the device is below or in the left and oriented away from the given trajectory. In the case, when the mobility device is above or in the right and oriented away from the given trajectory, the angular velocity of the left wheel must be higher than the angular velocity of the right wheel. The fuzzy logic is described in the control set shown Table 3.

Table 3. Fuzzy rules used for control

	Θ_H^-	Θ_M^-	Θ_Z	Θ_M^+	Θ_H^+
A_H^-	κ_K^M	κ_K^H	κ_K^H	κ_K^H	κ_K^H
	κ_D^M	κ_D^M	κ_D^-	κ_D^-	κ_D^-
A_M^-	κ_K^M	κ_K^M	κ_K^M	κ_K^H	κ_K^H
	κ_D^M	κ_D^Z	κ_D^Z	κ_D^Z	κ_D^-
A_Z	κ_K^-	κ_K^-	κ_K^M	κ_K^Z	κ_K^M
	κ_D^M	κ_D^Z	κ_D^M	κ_D^-	κ_D^-
A_M^+	κ_K^-	κ_K^Z	κ_K^Z	κ_K^Z	κ_K^M
	κ_D^H	κ_D^H	κ_D^M	κ_D^M	κ_D^-
A_H^+	κ_K^-	κ_K^-	κ_K^-	κ_K^M	κ_K^M
	κ_D^H	κ_D^H	κ_D^H	κ_D^H	κ_D^M

The fuzzy controller has been designed with two inputs (A – the shortest distance of the center to the given trajectory, $\Delta\Theta$ – angle between the trajectory and orientation line) and two outputs (κ_K - the angular velocity of the left side, κ_D - the angular velocity of the right side). The distance A is the first input described by 5 variables: A_H^- .high negative, A_M^- mean negative distances; D_Z - zero distance; A_M^+ mean positive, A_H^+ high positive distances. The angle $\Delta\Theta$ is the second input similarly described by another 5 variables: $\Delta\Theta_H^-$ .high negative, $\Delta\Theta_M^-$ mean negative angles; $\Delta\Theta_Z$ - zero angle; $\Delta\Theta_M^+$ average positive, $\Delta\Theta_H^+$ high positive angles. Two outputs κ_K and κ_D are described by 3 variables: $\kappa_{K,D}^Z$.- zero angular velocity of the wheels, $\kappa_{K,D}^M$ - mean angular velocity of the wheels, $\kappa_{K,D}^H$ - high angular velocity of the wheels.

4 Conclusions

Multimodal voice controlled assistive system for disabled people has been proposed. System allows recognize isolated commands together with some keywords. Important property of the proposed voice interface is that it works in permanent input mode and user did not need to press or release any key before speaking. To achieve this level of flexibility speech detection in long audio recordings algorithm was proposed. The algorithm is robust for a wide class of different noises and white range of SNRs. Experimental evaluation showed that using of speech detection algorithm allowed increase voice command recognition accuracy comparing with the recognition when detection wasn't used. Relative increase in the recognition accuracy was bigger in the case when the environment is noisy (noise type is additive and white). The increase of recognition accuracy was also observed in the acoustically complicated environments too. This observation allows us to make conclusion that recognition gains will be obtained using other sets of voice commands and will not be limited with the particular set implemented for the wheelchair control.

More convenient control of the assistive tools often requires better interpretation of recognized voice command. This is caused by the fact that some actions require the outcome which depends on the environment and other factors. For the wheelchair movement control intelligent control system was proposed.

The fuzzy logic based control strategy was deployed to execute the control, guiding an autonomous device along the indoor environments by using the robustness feature of the fuzzy controller design in the processing noisy and delayed data.

Acknowledgments. Parts of this work were done under the research project "Lietuviškų balso komandų atpažinimui orientuoto, multimodalinio išmaniųjų įrenginių asociatyvinio valdymo algoritmo sukūrimas ir modeliavimas", No.: 20101216-90 funded by EU SF project "Postdoctoral Fellowship Implementation in Lithuania" (VP1-3.1-ŠMM-01) of the Program of Human Resources Development Action Plan.

References

1. Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Voice interactive systems. In: Helal, A., Mokhtari, M., Abdulrazak, B. (eds.) The Engineering Handbook of Smart Technology for Aging, Disability and Independence, pp. 281–297 (2008)
2. Macek, T., Kleindienst, J., Krchal, J., Seredi, L.: Multi-modal telephony services in home Intelligent Environments. In: 3rd IET International Conference, pp. 404–410 (2007)
3. Rabiner, L.R., Sambur, M.R.: An Algorithm For Determining the Endpoints in Isolated Utterances. Bell System Tech J. (54), 297–315 (1975)

4. Ying, G.S., Mitchell, C.D., Jamieson, L.: Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement. In: Proc. of ICASSP 1993, pp. 732–735 (1993)
5. Hoyt, J., Wechsler, H.: Detection of Human Speech in Structured Noise. In: Proc. of ICASSP 1994, pp. 237–240 (1994)
6. Scheirer, E., Slaney, M.: Construction of Robust Multifeature Speech / Music Discriminator. In: Proc. of ICASSP 1997, pp. 1331–1334 (1997)
7. Rudzionis, A., Rudzionis, V.: Noisy speech detection and endpointing. In: Proc. of ISCA Workshop "Voice Operated Telecom Services", Ghent, Belgium, pp. 79–84 (2000)
8. Rudžionis, V., Maskeliūnas, R., Rudžionis, A.: Assistive Tools for the Motor-Handicapped People Using Speech Technologies: Lithuanian Case. In: Abramowicz, W., Maciaszek, L., Węcel, K. (eds.) BIS 2011 Workshops, Part 2. LNBIP, vol. 97, pp. 123–131. Springer, Heidelberg (2011)
9. Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Investigation of foreign languages models for Lithuanian speech recognition. Elektronika ir Elektrotechnika 3, 15–20 (2009)
10. Jurafsky, D., Martin, J.: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd edn. Prentice-Hall (2009)
11. Gokhan Ak, A., Cansever, G., Delibasi, A.: Robot Trajectory Tracking with Adaptive RBFNN-Based Fuzzy Sliding Mode Control. Information Technology and Control 40(2), 151–156 (2011)
12. Sun, D., Feng, G., Lam, C.M., Dong, H.: Orientation control of a differential mobile robot through wheel synchronization. IEEE/ASME Trans. on Mechatronics 10(3), 345–351 (2005)
13. Boquete, V., Garcia, R., Barea, R., Mazo, M.: Neural control of the movements of the wheelchair. Journal of Intelligent and Robotic Systems, 213–226 (1999)
14. Tang Shu-bo, Z., Yan, L., Lei, W.: Discrete trajectory tracking control of wheeled mobile robots. In: Proc. of the 2004 IEEE Int. Conf. on Robotics and Biometrics 2004, pp. 344–349 (2004)