Assistive Tools for the Motor-Handicapped People Using Speech Technologies: Lithuanian Case

Vytautas Rudžionis, Rytis Maskeliūnas, and Algimantas Rudžionis

Department of Informatics,
Vilnius university Kaunas faculty, Lithuania
Kaunas university of technology, Lithuania

Abstract. The paper presents analysis of the possibilities to use voice technologies for the partial integration of people with disabilities. The particular interest has been expressed to the motor-handicapped people. The special wheelchair with the voice command recognition capabilities has been designed. Evaluation of command's recognition accuracy shows high dependency on the proper detection of the utterance boundaries. The acoustic boundaries detection algorithm has been proposed. This algorithm allowed achieve high accuracy of the detection of acoustic events boundaries such as words or phrases even in the presense of high noise. The proper detection leads to the increased accuracy of voice commands recognition and the overall satisfaction of users.

Keywords: voice technology, voice command recognition, motor-handicapped people, acoustic events, detection of people speaking.

1 Introduction

People with disabilities meet barriers of all types. However, technology can help to lower many of them. Using computing technology tasks such as reading and writing documents, communicating with others, searching information on the internet, and controlling or even adapting the surrounding environment people with disabilities are capable of handling a wider range of activities independently. However typical input modes used in modern control and communication devices – keyboard input and visual output – are often not well suited for the disabled people. Voice technology is often the most preferable way for such people – to use speech recognition for recognizing commands as a substitute for keyboard or mouse based control and to use speech synthesis to read the content of computer screen as a substitute of typical screen reading by eyes. The ability to control the home is an essential aspect of independence and e-inclusion.

The implicit richness of human speech communication gives the user many degrees of freedom for control and input of various devices [1]. The speed of speech recognition also gives it a potential advantage over other input or control methods. 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

W. Abramowicz, L. Maciaszek, K. Węcel (Eds.): BIS 2011 Workshops, LNBIP 97, pp. 123–131, 2011.

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 can not benefit from one or another aspect of voice technology.

The barriers arising for the people with disabilities may be well shown on the example of the GUI development. One of the major advances of the human-machine interfaces in the recent decades was the advent of the graphical-user interfaces (GUI). It is hard to predict how much GUI was the reason of the massive computerization of the society but it is undoubtedly that GUI had significant impact. GUI popularized icons and the use of mouse for computer navigation which is generally extremely comfortable way to communicate with the computer for the majority of people. GUI is wonderful for the WYSIWIG (what you see is what you get) world, but is inconvenient for the people with various impairments such as visual, hearing loss or motor-handicapped persons. It is noted that prior to the GUI era [2] it was easier for impaired people to work with the computer using a Braille reader to assist them. In those days mainframes with character-based user interfaces were mainly used. System navigation was done using menus, tabbing, and function keys – all of which could be learned on a keyboard by the blind people. If the situation for the hearing impaired or the motor-handicapped people might look different in fact it is not a far better. And everyone can see the obvious reason for that: impaired or disabled people can't use one or more modalities that conforms the foundation of modern human-machine interfaces – monitor, keyboard or mouse. Other modalities are necessary in such situations. Well known fact is that speech is the preferable modality for the majority of the disabled people. If the person is mobility impaired, and cannot use their hands to move a mouse or type, issuing a voice commands is the natural solution. If people can't see, then applying text-to-speech technique to read the content is the most convenient solution. Voice technologies are the key element in the devices that are developed to satisfy the needs of many impaired people.

The success of the development of specialized tools for impaired people mainly relies on two factors: development of the voice technology being used as well as the knowledge of the special requirements of the disabled people. It should be emphasized that disabled people are especially valuable users from the point of view of voice technologists since due to the physical limitations they are ready to use even technologically restricted applications and not very well developed technologies that normal people often simply refuse to do.

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 [3]. 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.

Very important characteristic of voice based interfaces is the dependability of the phonetic, syntactic and lexical properties of the language spoken by the user. This means that it is impossible to move technologies developed for the recognition of one language for the recognition of another automatically. Some sort of adaptation would be necessary. Since major developers of speech technologies aren't particularly interested in less spoken languages such as Lithuanian the need for adaptation in such cases is even more important. One of the possible solutions for some class of applications is the adaptation of foreign language based speech engines via the selection of proper phonetic transcriptions. In our previous studies the advantages of such method and its possible uses were established [4, 5]. But the success of voice based interfaces significantly depends on the proper detection of speech in long recordings: such devices typically work in the continuous recording mode while only very small part of recording contains useful information. At the same time often users need to operate in the noisy environments what makes the problem of detection of the speech boundaries not trivial.

Further the paper is organized as follows: the second part presents voice controlled wheelchair with the possibilities to implement multimodal control opportunities. The third part presents acoustic events detection algorithm and it’s evaluation in different acoustic environments. And finally achieved results and further prospects are summarized.

2 Wheelchair with Multimodal Control Possibilities

Wheelchair is one of the main assistive tools used by the motor-handicapped people in their daily lives. Wheelchair provides the opportunity to overcome the main limitation of this type of people – the inability to move independently. There were many different types of wheelchairs proposed and used in practice. The simplest type is manually controlled and relies on the assistance provided by the third person: the supporting person moves the wheelchair to the place were disabled person wish to be. Such wheelchairs has obvious disadvantage: they need the assistant person to be in a close proximity to the disabled person to be called when necessary. Often the need to ask for an external help arouses psychological problems to the disabled person since the necessity to ask for a help each time emphasizes dependency from other people. The reliance on own hands and the possibility to use own force to move such type of

wheelchairs isn't possible for all people with motoric system disabilities. Another type of wheelchairs uses the electric motor and battery to move the wheelchair. The user typically controls the equipment with the small keyboard or even some kind of joystick type device. Being significantly more convenient than pushcart wheelchairs such kind of tools has also several drawbacks. One of them is the dependency from battery: the more it is used the more likely it will require loading. The loading means that wheelchair will be unavailable for the disabled person for a while. Another drawback is that it is difficult and inconvenient to use it in the small spaces such as living rooms, corridors, etc. This means that it would be highly desirable to free the person from the necessity to move when doing such tasks as switching on/off lights, turning on/off radio, etc. And for some types of diseases even the using of keyboard or joystick to control the wheelchair is problematic.

There were successful attempts to do wheelchairs using voice commands as the mode for the control. Such wheelchairs typically has embedded voice command recognition and control unit designed to recognize and process pre-specified set of commands. From a Lithuanian speaker perspective it is very important that such wheelchairs recognizes only English (or some other language) commands and it is difficult to embed the recognition of Lithuanian commands.

These considerations suggested us to propose client-server based architecture for wheelchair control: the user is provided with the PDA type device which serves as the recorder and does some initial processing of speech signal and transmits it to the server. Server runs speech recognition engine, receives voice commands, recognizes them and makes appropriate turns. The PDA client and server are linked using Bluetooth or wireless connection. Such approach enables to expand human-computer interaction with the additional modes easier and in a more flexible way in the future. The Fig. 1 shows the principal schema of client-server structure of wheelchair voice command based control system:

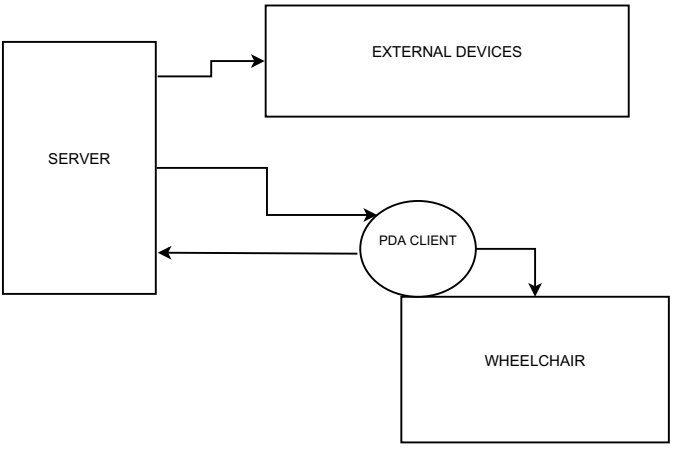

Fig. 1. Client-server architecture of voice command based control system for wheelchair and additional devices

The Fig. 2 shows the wheelchair equipped with the voice commands recording and processing equipment.

Fig. 2. Wheelchair with equipment for the recognition of voice commands

The main advantage of such approach is it flexibility: since nowadays houses are equipped with a set of household appliances ranging from simple lights to feature-rich hi-fi systems, DVD players and TV sets then it is possible to use voice command recognition system to control this big variety of home appliances not only the wheelchair. In this case speech server has controller with executive unit connected with selected appliances. Another advantage is the possibility to implement other than voice commands control possibilities or to use them as the additional channel for the commands. In this case it is possible to organize multimodal wheelchair control system.

From the developer's point of view the main advantage is the possibility to implement wide range of voice commands recognition algorithms. It also includes the possibility to use third party or foreign language speech recognition engines and to implement the experience gathered adapting foreign language speech recognition engines for the recognition of Lithuanian voice commands. It has been shown earlier that such adaptation enables to achieve voice command recognition accuracy necessary for the commercial applications (95% and above) while applying only limited resources. To achieve this goal in open and noisy room conditions some additional requirements should be met. One of them is the proper detection of acoustic boundaries in the noisy recordings: since it is more convenient for user to record audio signal continuously and to recognize command just after the utterance was finished rather than to press a key and begin to talk we need to detect reliably the boundaries of acoustic events (words, phrases, utterances). The reliability should be maintained even if the environment is noisy. The next chapter describes proposed method for the detection of the acoustic events in long or continuous recordings.

3 Detection of Acoustic Events in Long Utterances

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 the area of speech processing. It is not surprisingly that a lot of activities were devoted to solve this problem. Various

algorithms proposed for the detection of speech and segmentation of spoken utterances are presented in [6-9] and others sources. Most of the algorithms exploit such spoken signal properties such as the articulatory movement's features or the differences between the actual signal spectrum and the spectrum prediction using its first or second order regression. 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 various.

Many methods also are based on the signal energy changes as one of the factors, which reflect best the acoustic changes in a speech signal.

In this study detection of acoustic boundaries is important as a template approach for the detection of phoneme boundaries using visual features: the quality of the detection using visual features should be compared with the results achieved using the detection based on articulatory features. Of course it could be possible to use manual segmentation of spoken signal but in this case the complexity of the study will grow enormously. The combined use of the acoustic and visual segmentation (e.g. for the improved recognition of speech or the detection who from a group of people is speaking) is still the future goal and isn't covered by this study. But we need to use robust algorithm to segment the spoken signal to use automatically generated boundaries of acoustic events as a templates for visual features. In this study we used proper algorithm for acoustic events detection which showed to be accurate and robust enough for the segmentation of spoken speech in previous study [10].

In this study we used slightly modified algorithm. The essence and the background could be explained as follows. It is well known that the speech is non-stationary process over longer time spans. At the same time speech could be considered as a quasi-stationary process over shorter time periods (a time frame is no longer than 30 ms though the exact duration of the stationarity depends on the phonetic content of a signal). Most algorithms count on the periods, where the statistical properties of stochastic process change moderately. Since speech signal could be described as a process with time varying frequency, the properties of a speech signal in the frequency (or spectral) domain is rather informative.

At the initial stage the logarithmic spectrum was derived using 8-10 msec step. Experimental evidence proved that the spectrum based on the recursive IIR filter bank is more robust to fluctuations of the spectrum properties on the adjacent speech frames. These vectors were used to construct the likelihood function of the changes in spoken speech. The changes in the likelihood function values enables to capture such highly indicative acoustic event features as the articulatory movements. Comparing with the method described in [10] we used shorter period for the integration of the likelihood function. The integration is necessary since it allows obtaining smoother likelihood function and helps to avoid the random type of fluctuations which are characteristic for the changes in spectral properties of many phonetic units. The function f_{SR} was used to perform smoothing by converting the sequences of the likelihood function parameters r_t to smoothed sequence of the parameters ρ_t:

$$\rho_t^{SR} = f^{SR}(r_t, \tau, \nu) \quad (1),$$

were the parameters τ and v defines the smoothing duration and the smoothing type. The sequence ρ_t is used for the detection of the boundaries of acoustic events and it is

called an acoustic events response (AER). The examples of AER are presented in the Figure 1, which consists of the oscilogram of the original and the differentiated syllable, the spectrogram of the same syllable and the AER curve. 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 a peak of the AER the higher is the likelihood value of the boundary between the different acoustic units.

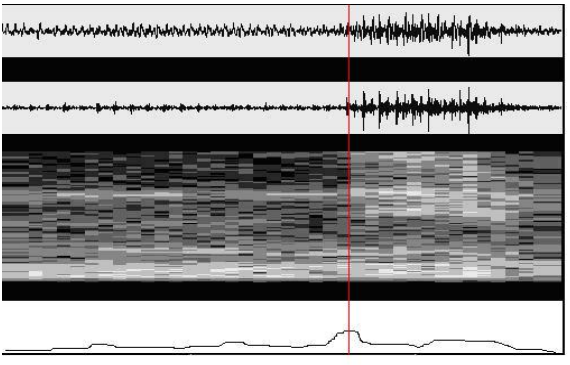

Fig. 3. The acoustic events likelihood function within an isolated word

The efficiency of basic algorithm for the detection of endpoints of spoken utterance in noisy recordings has been investigated in [10]. Then was showed that the algorithm allows detect the boundaries of words in noisy recordings with high accuracy even when the SNR ratio falls down to -30 dB. In this study we analyzed the efficiency of the modified algorithm for fixing the boundaries of phonemes in diphones composed of the nasals (m,n) and six vowels (a,o,e,ee,u,i). The noise has been added by carrying out the analysis with different SNR levels. An expert labeler analyzed diphone recordings and checked the starting and ending positions of the diphone as well as the boundary between a nasal and a vowel. Those labels were considered as the references for the automatic algorithm. In case the boundary detected by the algorithm felt into the range $x \pm 20$ msec of the manually detected boundary, the result was considered as correct. Otherwise an error was fixed. Two types of errors were analyzed: false alarm P_{fa} and false rejection P_{fr}. Table 1 presents the averaged errors per speaker for different SNR values.

Table 1. Average number of false alarm and false rejection errors detecting the boundaries of phonemes in diphones nasal-vowel

SNR, dB	Average number of errors per speaker	
	P_{fa}	P_{fr}
12	1.5	0
6	2.2	0
0	8.3	4.9

It could be concluded that the algorithm based only on audio information reliably detects the boundaries of phonemes in diphones when the noise level is low or moderate. When the noise level substantially increases the algorithm produces relatively great number of errors (8.5 false alarms and 5.4 false rejections per speaker; having in mind that each speaker pronounced 24 diphones). In the presence of a noise the information of lip movements should form a complementary source for improving the detection of acoustic event boundaries.

4 Conclusions

The wheelchair with voice command based control was developed. The wheelchair control interface has embedded capabilities to be expanded to the multimodal interface implementing several modes of human-computer interaction.

The recognition of voice commands is performed using client-service architecture: the recognition engine is implemented at the server side while client serves only as speech recorder and transmitter to the server. Such approach enables to achieve higher recognition accuracy exploiting higher resources of the server side computer and consequently more sophisticated algorithms. This approach enables to expand the human-computer interaction mode with more modalities in the future.

The user's satisfaction is affected by the commands recognition accuracy. The recognition accuracy is affected by many factors. One of affecting factors is the proper detection of the spoken utterance boundaries. The algorithm for the detection of acoustic events boundaries in long and noisy recordings has been proposed. The algorithm enables to detect the boundaries of acoustic events even when SNR level goes down to 3-6 dB. The proper detection of boundaries should lead to the overall increase in the voice commands recognition accuracy.

Acknowledgements. Parts of this work are part of the research project "Lietuviškų balso komandų atpažinimui orientuoto, multimodalinio išmaniųjų įrenginių asociatyvinio valdymo algoritmo sukūrimas ir modeliavimas", No.: 20101216-90 funded by European Union Structural Funds project "Postdoctoral Fellowship Implementation in Lithuania" within the framework of the Measure for Enhancing Mobility of Scholars and Other Researchers and the Promotion of Student Research (VP1-3.1-ŠMM-01) of the Program of Human Resources Development Action Plan.

References

1. Hawley, M., Green, P., Enderby, P., Cunnigham, S., Moore, R.: Speech technology for e-inclusion of people with physical disabilities and disordered speech. In: Proc. of Interspeech 2005, Lisbon, Portugal (2005)
2. Rogoff, B., Goodman Turkanis, C., Bartlett, L.: Learning together: Chikdren and adults in school community. Oxford University Press, New York (2001)
3. Macek, T., Kleindienst, J., Krchal, J., Seredi, L.: Multi-modal telephony services in hometal Intelligent Environments. In: 3rd IET International Conference, pp. 404–410 (2007)

4. Maskeliunas, R., Rudzionis, A., Rudzionis, V.: Advances on the Use of the Foreign Language Recognizer. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) Second COST 2102. LNCS, vol. 5967, pp. 217–224. Springer, Heidelberg (2010)
5. Maskeliunas, R., Rudzionis, A., Ratkevicius, K., Rudzionis, V.: Investigation of Foreign Languages Models for Lithuanian Speech Recognition. Electronics and Electrical Engineering – Kaunas: Technologija 3(91), 37–42 (2009)
6. Rabiner, L.R., Sambur, M.R.: An Algorithm For Determining the Endpoints in Isolated Utterances. Bell System Tech. J. 54, 297–315 (1975)
7. 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)
8. Hoyt, J., Wechsler, H.: Detection of Human Speech in Structured Noise. In: Proc. of ICASSP 1994, pp. 237–240 (1994)
9. Scheirer, E., Slaney, M.: Construction of Robust Multifeature Speech / Music Discriminator. In: Proc. of ICASSP 1997, pp. 1331–1334 (1997)
10. Rudzionis, A., Rudzionis, V.: Noisy speech detection and endpointing. In: Proc. of ISCA Workshop “Voice Operated Telecom Services”, Ghent, Belgium, pp. 79–84 (May 2000)