Voice Technology Applied for Building a Prototype Smart Room

Josef Chaloupka, Jan Nouza, Jindrich Zdansky, Petr Cerva, Jan Silovsky, and Martin Kroul

Institute of Information Technology, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic
{josef.chaloupka,jan.nouza,jindrich.zdansky,petr.cerva, jan.silovsky,martin.kroul}@tul.cz
http://itakura.kes.tul.cz

Abstract. This contribution is about a system called VoiCenter that allows motor-handicapped people to control PC and standard electric and electronics devices (lights, heating, climate control, electric blinds, TV, radio, DVD, HI-FI etc.) in their homes. The PC and the devices are possible to be controlled with the help of simple voice commands. The wireless connection between PC and the devices was used in this system. This is good for fast installation of this complex system in homes. We have used our own speech recognition and control computer system that is described in this paper too.

Keywords: SmartRoom, Intelligent Home for a Motor-Handicapped Person, Voice Control of PC, Electronic and Electric Devices.

1 Introduction

The real results of many-years research in the area of speech processing and recognition are programs which allow to convert a spoken language to either text or to PC commands in the real time. These applications depend on a particular national language for which the system is developed. The commercial systems especially for voice dictation exist for some world languages - English, French, German or Japanese. We are focused on Czech language because similar commercial systems don't exist till today (2007). The reason is that the Czech language is relatively complicated and it belongs to the inflected languages, therefore the word vocabulary is multiple larger than for example the English vocabulary. We have developed and created several systems with our own speech recognition toolkit in our laboratory SpeechLab in the Technical University of Liberec. The first useful Czech system was the automatic information phone system InfoCity [1] for human - computer dialog via phone (1999). The prototype of Czech dictation system with a vocabulary of 800 000 Czech words was the second system [2] and we have been creating the system for automatic broadcast news transcription (since 2004). The next system was the program MyVoice

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 104–111, 2009.

(2005) for voice control of a PC [3]. Program MyVoice is developed for motor-handicapped people for whom it isn't possible to use mouse or keyboard and voice is the only possibility how to control any programs in PC. Several tens of handicapped people are using this program at present. We have modified the program MyVoice according to wishes of handicapped people but these people are telling us very often that they have different standard electronic and electric devices in their homes and these motor-handicapped people can not use and control these devices. Therefore the goal of this work was to develop and create some complex system for voice control of PC together with home electric and electronic devices. The modified program MyVoice is the "heart" of this system - we call it VoiCenter.

2 The Voice Control System VoiCenter

This system VoiCenter was developed mainly for motor-handicapped people for controlling of PC and electric and electronics devices in their homes. Several different ways exist at present how to control PC without hands [7], but the voice is the most natural way of them. But the premise is that they can speak. The system VoiCenter consists of several subsystems (modules). They are the following: automatic speaker identification and verification module (ASI/V), automatic speech recognition (ASR) module together with control PC unit - modified program MyVoice, unit for controlling outside devices, and text to speech (TTS) synthesis module, see fig. 1. Single modules and their utilization are described in the following chapter.

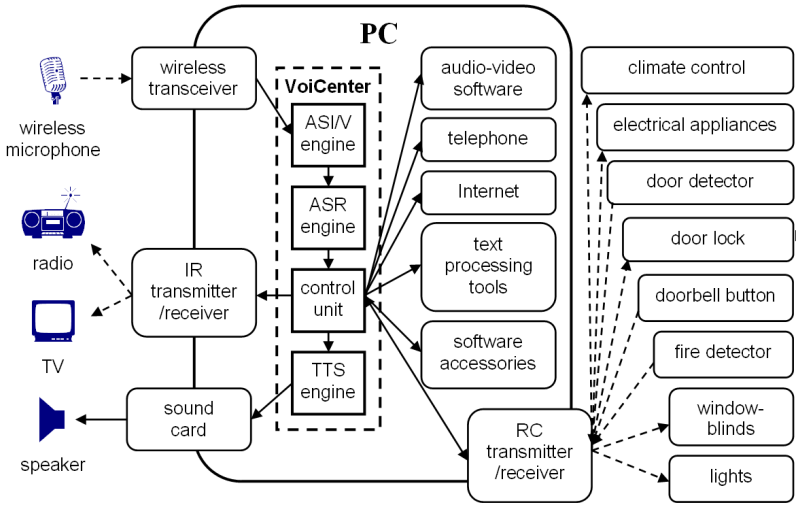

Fig. 1. The system VoiCenter for controlling electric and electronic devices in home of a handicapped person

2.1 Voice Control for PC - Program MyVoice

The core of this program is the unit for automatic speech recognition that was created in our laboratory [3]. This unit works with vocabularies that contain hundreds to thousands of words. It is possible to change a content of that vocabulary according to the settings of the program MyVoice for controlling of single programs in the PC. The recognition system is based on HMM's of single Czech phonemes. The HMM's were trained on several tens of hours of Czech utterances of hundreds of human speakers. Therefore speech recognition is speaker independent (but it is language dependent) and it is relatively easy to add a new word to the vocabulary. Because it is necessary to have a phonetic transcription of this word (or word connection), the phonetic transcription subsystem is a part of the program MyVoice. It is necessary to manually improve phonetic transcription (mainly for foreign and special words) sometimes but it happens only in special cases. Several tens of word groups are for different actions in the program MyVoice and it is possible to switch between single groups by voice commands, for example: One group is for movement in operation system desktop. The voice commands in this group are the following: "Left", "Right", "Down", "Up" and "Take it" - these are equivalent to keyboard keys. It is very easy to select some icon of a program in the desktop and to run the selected program.

If the computer user wants to write text and some text editor was run, then the next group is used for working with text. It is possible to write the text with the help of letters. Where single letters are dictated like: "A" or "Alpha", "B" or "Bravo", etc. or we use for dictation some words (as much as several thousands) that are included in this group. Another group is used for mouse arrow movements. The commands here are like: "Left 10", "Right 50", "Up 200", "To middle", "Click", etc. The principle of the program MyVoice is the following: It connects some virtual computer system action(s) (from mouse or keyboard) to voice commands. New words (voice commands) and their system actions are set in the configuration window in the program MyVoice. If the voice command is recognized, the program MyVoice sends virtual action(s) to the active program (text editor, web browser, game ...). It is possible to pronounce a name of any key in keyboard and it is possible to simulate every mouse or keyboard action in the program MyVoice, therefore we can do every action with any program in PC without using our hands.

What is possible?

1. Run program Calculator, enter numbers and operation with the numbers, show result and use it in some another program if we want it.
2. Run some text editor (Notepad, MS Word or ...), dictate text letter by letter or with the help of words or word connections, format text with select parameters (size of text, font, line spacing, etc.), load and save a text, print a text.
3. Work with Internet, browse any web-pages, look for some information in web-pages, chat with other people and write or read emails.

4. Paint some pictures in different image programs, browse images, photos and videos, play games if we know witch keys or mouse actions are used for playing, play music, etc.
5. If we have in our computer TV or radio card, it is possible to watch TV programs or hear to the radio broadcasting, read a teletext, choose TV and radio stations. If we have a phone card in the PC too, then it is possible to dial phone numbers, talk with friends or with doctors, policemen or firemen (if we need it). These actions are all without using hands, only by simple voice commands.
6. Control electric and electronics devices in the home with the help of some special software and interface that allows communication between PC and the device.

2.2 Speaker Adaptation

We have used speaker independent (SI) HMM's in our speech recognition system but sometimes it is good to use speaker adapted (SA) HMM's if the speaker has some speech defect or if we want to have better speech recognition result. The speaker adaptation module is employed to improve the speech recognition accuracy for individual users of the software. It is very useful to incorporate such a module to the recognition system for handicapped persons because their physical handicap is often connected with a various speech defect (like for people with dysarthria for example).For adaptation, we use the combination of MAP [4] and MLLR [5] methods and gender dependent (GD) models as prior sources: the given GD model (male or female) is adapted by the MLLR method at first and then the created model is used for the MAP based adaptation. The advantage of this approach consists in the fact that also the models not seen in the adaptation data can be adapted (by MLLR) while the parameters of the models with a lot of adaptation data can converge to the values of the theoretically best speaker dependent model (due to the MAP method). The results of performed experiments (see [3] for details) have shown us that the recognition accuracy in the program MyVoice for a quadriplegic girl with a moderate speech defect can be improved by adaptation from the level of 83% on the level of 93% which is not so far from the accuracy for persons with standard pronunciation (97%).

2.3 Practical Experience with Program MyVoice

Practical tests show that it is possible to do any action in PC which is possible to be done with a classical input PC devices - keyboard and mouse with the help of the program MyVoice. The voice command recognition is relatively reliable because it is based on well-tried and many years developed technologies. A user of this program can not narrate any word because all program is based on speaker independent HMM's of single phonemes but a necessary condition is that a human speaker pronounces the voice commands clearly and correctly. It is possible to modify the vocabulary very easily in the configuration program

window. The actual group of words is shown in the main program window, therefore it isn't necessary to remember all voice commands. The program is running in the background of active programs and it occupies only a small place on the screen. The vocabulary and recognized words are shown only in this place. The program MyVoice can run well on a PC that has more than 500MHz and the memory larger than 128MB. Then we need a standard sound card and microphone with headphones, and the program is useful only in PC with operation system Windows 2000 and higher. It is possible to order the most frequently used words or sentences by one simple voice command if we dictate some text in the text editor, email client etc. The user can create and use the collection of voice commands for the most frequently used programs like MS Word or for some web browser. It markedly makes working with these programs easy because we can connect several system actions with a single voice command, for example: a single voice command runs some program, opens a new document, sets-up size and fonts, and switches a new word group with the voice commands for this program to the main window of the program MyVoice. The control of the computer mouse is somewhat more difficult with the voice commands then if the user uses the hands but it is possible with a batch of voice commands to move mouse arrow to the exact position in the screen and to simulate a mouse click or double click. More than 50 people use the program MyVoice at present and most of them are motor-handicapped people which told us about their experience with this program and we are improving the program according to their requirements.

2.4 Speaker Verification and Identification Module in VoiCenter

The speaker verification (ASI/V) module is used to secure authorized access to the system VoiCenter only to certain persons or to constrain a command set depending on the speaker identity. This module can be disabled if a handicapped person is alone at home but if not it is reasonable to incorporate such a mechanism since for example children should not have the possibility to tell as much commands as their parents (e.g. cancel the fire alarm). Based on the role assigned to the identified user, a voiced command is accepted and executed or rejected. As the most of state-of-the-art speaker verification systems, our SV module is based on modeling short-time cepstral features by Gaussian mixture models (GMM) and uses a GMM-UBM approach [6]. As well as for speech recognition, MFCC features were used with the difference that c_0, delta and delta-delta features were excluded. Thus feature vector was formed from 12 MFCCs. Speaker verification decision about the claimant identity is based on the log-likelihood ratio test. Claimed identity is accepted if the following condition holds.

$$\frac{1}{T}\left(P\left(X|\lambda^{s}\right) - P\left(X|\lambda^{UBM}\right)\right) > \Theta \tag{1}$$

where λ^s represents a GMM of a speaker s, λ^{UBM} represents a universal background model, $P\left(X|\lambda\right)$ is the log-likelihood function for a GMM λ, T is the

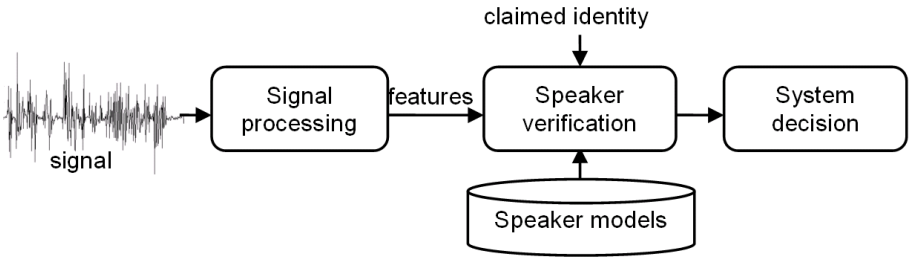

Fig. 2. Principle of speaker verification and identification module in VoiCenter

number of observation used for likelihood evaluation, and finally Θ is a verification threshold.

3 Modules and Interfaces for Controlling of Electric and Electronics Devices in the Homes

At present, a lot of different modules, system and interfaces from some companies for controlling of electric and electronics devices in the homes exist. We wanted to have wireless connections between computers and these devices because the technical installation of such a system is more easier. We have chosen the wireless modules from the Czech company Jablotron. The reason for this choice was that this company is very near to our University therefore the communication with this company was easier for us. We have used wireless (radio) receivers that have two relays which can be controlled. These modules were used for controlling the lights, window-blinds, electric heating and ventilator.

Control of some other electric devices like electric cooker or electric tea-kettle doesn't have sense for motor-handicapped people because they don't have chance to manipulate with these devices. We have got some information how it is possible to use their receiver and transmitter from company Jablotron therefore we have created our own universal transmitter-receiver interface that communicates between PC and wireless modules. This interface can control 20 electric devices in home and it can receive signal from 4 wireless sensors - we have wireless doorbell button, fire detector, door detector and motion detector. The state change of these sensors are shown in the system VoiCenter and the information mesage (audio signal) is sent over sound card to an electronic speaker if the flag output sound is set in the VoiCenter configuration window.

The Text-To-Speech (TTS) synthesis module is used for conversion of information message (sentence) to audio signal. A lot of home electronic devices (TV, radio, HI-FI) have infrared remote controller. Therefore we have used universal infrared remote controller that is connected with PC so that it is possible to control these devices by voice commands. We have a special word group for each infrared controlled electronic device in our system therefore all actions (turns the device on and off, volume increase/decrease, menu, teletext...) in these devices are controlled from PC.

4 Text-To-Speech Synthesis Module

The Text-To-Speech conversion module is used to provide computer a possibility to respond to user's commands. It enables a backward communication in computer-user direction and is a comfortable way how to deliver requested information to the user. For user it is convenient, that he or she doesn't have to look onto monitor after every command to verify if it was recognized properly. Thus he or she can move freely and is not forced to stay in one place. Various kinds of information can be preferably transmitted by voice, e.g. state of the system, command confirmation or alerts. With properly composed dialogue system, monitor becomes unnecessary. Our TTS conversion module is based on concatenation of triphone speech units, with intonation contour modification using TD-PSOLA method. This is today's very common approach. Scheme of the module is in fig.3. Input text is preprocessed first.

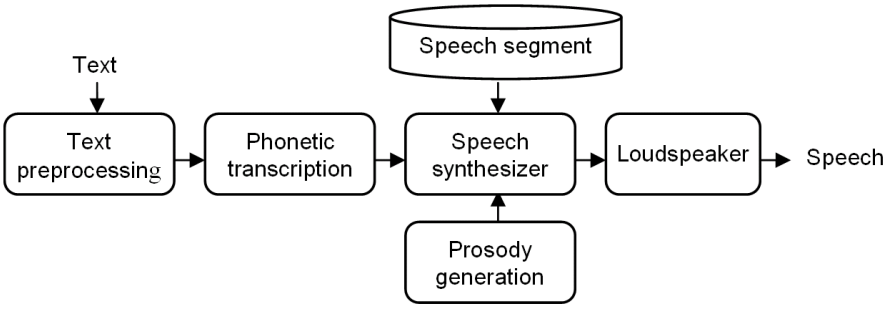

Fig. 3. Scheme of Text-To-Speech (TTS) module

Abbreviations, numbers, web and e-mail addresses are translated to their spoken form (sequence of words), non-printable and unwanted symbols (like *, $ or #) are filtered. Phonetic transcription is performed then - the text is transcribed to sequence of phonemes and sent to a synthesizer. The synthesizer searches for appropriate triphones in a database, concatenates them and applies prosody (intonation, accent and rhythm). Finally, speech signal of the created utterance is sent to a loudspeaker.

5 Conclusion

The system VoiCenter can help motor-handicapped people in many ways. Above all they can work with PC, write and read texts and emails. It is a new element in their lives because they can communicate with some other people from the world or they can start to study.

The work with web browser is relatively easy if they have internet connection, therefore handicapped people have a new possibility how to obtain new information from different web-pages, news, etc. But the novelty in this system compared

to the program MyVoice is that it is possible to control a lot of standard home electronics and electric devices from the PC by simple voice commands, therefore they don't need some special cards in PC and they can control such things like heating, climate control, lights etc. so that they are more free in their decisions.

The system VoiCenter is possible to be used in the homes of not-handicapped people too mainly for controlling of electric and electronics devices, but it is better to use computer mouse and keyboard for controlling the PC programs for them. The system VoiCenter is installed in our laboratory at present, but we would like to install this system in the home of some handicapped person in the near future.

Acknowledgments

The research was supported partly by the Grant Agency of the Czech Academy of Sciences (grant no.1QS108040569) and partly by the internal grant provided by the Faculty of Mechatronics at the Technical University of Liberec (grant no. FM-IG/2008/ITE-01).

References

1. Nouza, J., Holada, M.: A City Information System Operating over the Telephone. In: Proc. of IVTTA 1998 Workshop, Torino, Italy, vol. 1, pp. 141–144 (September 1998)
2. Nouza, J., Nouza, T.: A Voice Dictation System for a Million-Word Czech Vocabulary. In: Proc. of ICCCT 2004, Austin, USA, pp. 149–152 (August 2004) ISBN 980-6560-17-5
3. Nouza, J., Nouza, T., Éerva, P.: A Multi-Functional Voice-Control Aid for Disabled Persons. In: Specom 2005, Patras, Greece, pp. 715–718 (2005) ISBN 5-7452-0110-x
4. Gauvain, J.C., Lee, L.H.: Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Trans. SAP 2, 291–298 (1994)
5. Gales, M.J.F., Woodland, P.C.: Mean and Variance Adaptation Within the MLLR Framework. Computer Speech and Language 10, 249–264 (1996)
6. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
7. Evans, D.G., Drew, R.: Blenkhorn. P.: Controlling Mouse Pointer Position Using a Infrared Head-Operated Joystick. IEEE Transaction on Rehabilitation Engineering 8(1) (2000)