



Design and Development of Voice Controlled Aids for Motor-Handicapped Persons

Petr Cerva, Jan Nouza

SpeechLab, Institute of Information Technology and Electronics
 Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic
 {petr.cerva, jan.nouza}@vslib.cz

Abstract

In this paper we present two voice-operated systems that have been designed for Czech motor-handicapped people to allow them full access to computers and computer based services. The programs, which are named MyVoice and MyDictate, are complementary in their functions. Both employ ASR engines developed in our lab. The former is used primarily as a mid-size-vocabulary (up to 10K words) voice commander for PC programs and PC-controlled home devices, the latter allows for very-large-vocabulary dictation (with more than 500K words). They are designed to cooperate and thus allow an entirely hands-free access to any computer application, including text typing, e-mail exchange, Internet browsing or handling telephone calls as well as for controlling external home devices such a TV/radio sets or air-conditioning.

Index Terms: voice control, text dictation, speech recognition, handicapped persons, voice aids

1. Introduction

Recent progress in speech research together with faster and cheaper computers have made voice technology mature for practical use in areas like voice control of a PC or text dictation. Such programs already exist for major languages like English, French, German, or Japanese. Unfortunately, there are many other languages for which such tools have not been available and the chance they could be available soon is not very high. The reasons are twofold: a) minor languages offer much smaller market for sale and b) some of the tongues proved to be significantly more complex for automatic speech processing, when compared e.g. to English. In case of Czech, it is its inflective nature that makes speech recognition difficult mainly because most lexicon items can appear in many different forms, which are derived from their base-forms with respect to linguistic and grammatical context.

Nevertheless, there is a strong demand for computer systems that would accept input in spoken Czech. This demand is especially urgent in the case of motor-handicapped people, mainly because they are aware of the advantages that such systems suppose for the disabled persons who are already using them in other languages.

Two years ago we accepted the challenge and started to develop a program that could act as an interpreter of voice commands into standardized computer actions, namely virtual key-strokes and their sequences, virtual mouse movements and clicks and system messages sent to various software and hardware units. As a result, the program named MyVoice [1] was completed in 2005 and since that time, several tens of handicapped users have learned to use it. It allows them to control any application installed on their computers entirely by voice. Any program running under Microsoft Windows

OS can be started and controlled by voice commands imitating key-presses and mouse actions. In this way one can utilize an Internet browser, exchange e-mails, draw pictures, listen to music, or type text documents.

The typing of documents, however, has been limited in its practical use. Because the ASR engine employed for MyVoice was developed mainly with regards to robust performance and low-cost hardware, the largest vocabulary that MyVoice could manage was only 10 thousands most frequent words. In Czech language, that lexicon size covered just 65 % of the words encountered in common texts. All out-of-vocabulary items had to be spelled letter by letter. To help the target users, we decided to develop an additional tool that would make dictation more comfortable. Detailed analysis of large text corpora showed that the lexicon must have at least 500,000 items to reach 99 % coverage of spoken Czech. To manage such a large lexicon in real time and, at the same time, to allow for easy correction of wrongly recognized or phonetically ambiguous items, the dictation tool (later named MyDictate) was designed to work with discrete-speech input. It can be used as a standalone program or as an extension to the previously developed MyVoice.

In the following sections, we describe the ideas used in the design of both programs. In section 2, we describe their common platform and explain the main features of both ASR engines, each developed for a slightly different task. In next sections, MyVoice and MyDictate tools are presented. Section 5 then deals with evaluation tests performed for both systems. Finally, in section 6 we present some conclusions and propose several future work guidelines.

2. Voice Control vs. Unlimited Text Dictation

When developing both the tools, we had to keep in mind that most of the target PC users would have no chance to use a keyboard or mouse. It means that for every desired action there must be a way to accomplish it only by voice. In principle, the solution is rather simple: all keyboard and mouse actions must be replaced by voice input. However, we must take into account that the real effect of a single key press or a mouse click highly depends on context. It may start a new program, it may control the already running one, it may add a letter or a word into a text box or typed document, or – in a special case – it may have an impact on the voice recognition tool itself. All these different actions and situations are depicted in Fig. 1 and described in the following section.

2.1. Common platform

Our tools are based on a common recognition engine that process the speech signal from a microphone and translates it into events. These can allow namely for:

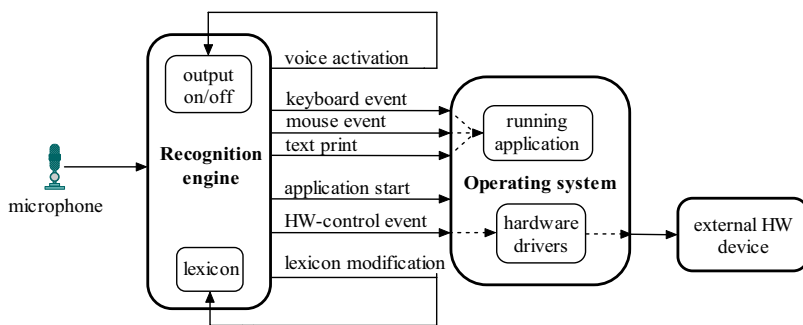


Figure 1: Scheme of speech recognition platform designed for voice control (MyVoice) as well as text dictation (MyDictate)

- activation or deactivation of the output from the engine to allow for a sleeping mode
- modification (swap or update) of the current lexicon
- simulation of key-press and mouse-event actions (and their combinations) to control running applications
- print of a text (according to the given local settings) into an active text window,
- launch of a program (together with its arguments)
- sending of messages to control external devices connected to the given PC

This platform is common for both the tools, but each of them employs only a selected part of the available actions and features, as they were designed for carrying out two slightly different types of tasks.

2.2. Two different tasks

On the one hand, the MyVoice tool accomplishes a first group of tasks which consist in starting programs, selecting items from menus or inputting some parameters in form of short sequence of letters and digits, using the mouse to move the cursor or objects, or to navigate to links. Many of these actions are hard to be cancelled or it is not easy to undo them. That is why this type of actions must be supported by very robust voice recognizer. This can be accomplished by keeping the list of currently applicable commands as low as possible. In order to do so, we adopted the good practice of grouping the commands into sets of related items that can be correctly discriminated from the ASR point of view and easily remembered by the user.

On the other hand, MyDictate was created for text dictation, which is a slightly different type of PC application. In this case, the active vocabulary must be very large (more than one hundred thousands items). This vocabulary changes only occasionally when the user wants to add a new word. The dictation tool must be designed to cope with this large load and it must take into account that the spoken text can be misrecognised and misinterpreted due to various reasons, like recognition errors, ambiguous items or alternate spelling, among others. Thus, MyDictate had to be equipped with features that allowed for easy and immediate correction of input text, selection from several acoustically similar (or even same) word candidates as well as any arbitrary editing of the already printed text.

For both the tasks, there must be an option to switch off the output of the recognizer temporarily, i.e. to simulate a sleeping mode. This is important in situations like when the user wants to speak to another person, etc.

2.3. Technical description of ASR engines

Both the engines share the same front-end and they differ only in the decoding module. The currently used front-end includes routines that control the sound-card and process an 8 kHz sampled signal from the microphone. Every 10 ms, a 25-ms long signal frame is parameterized to get 39 mel-frequency cepstral coefficients (MFCC) as

well as log energy, that is employed only for detecting starting and ending points of speech. The acoustic inventory is comprised of 40 Czech phonemes [2] and 8 types of noises (like breath). Each of these units is modeled by 3-state 64-mixture HMMs trained on 40-hours of Czech recordings.

The decoder employed in MyVoice has been optimized for frequently changing mid-size lexicons. A higher level of performance, which is crucial for robust recognition of executive commands, has been achieved by including multiple pronunciation variants for most lexicon items and rather conservative pruning strategy. On contrary, the decoder in MyDictate was designed to cope with very large lexicons containing up to 1 million items. To accomplish its real-time performance, a fast tree-search based decoder was designed to take into account also word frequencies and word similarities. This decoder outputs a list of candidates, from which the top ten are displayed to the user. In the case of a minor recognition error or lexical ambiguity, the user can pick up the desired word from the list by uttering a single command.

3. MyVoice - design and functions

The system MyVoice (as well as MyDictate) was designed to run under Microsoft Windows 2000 or XP, because these a) are widely used by Czech users and b) have an improved support of Czech characters. For users, the most important part of the software is a tiny main program window (Fig. 2), which is always on the top of all windows on the screen.

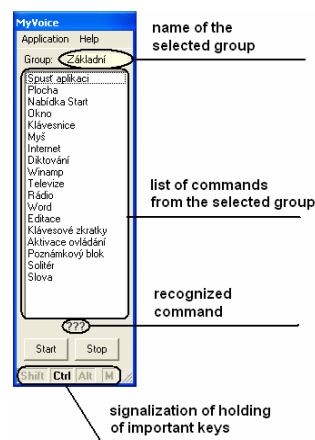


Figure 2: MyVoice's main window

Its size can be modified, but in principle it should occupy minimum space so that it does not hinder the visibility of the application that is currently being controlled. This window is divided into several parts (see Fig. 2) which provides the user with different type of visual feedback information.

3.1. Commands and command groups

The recent version of MyVoice contains 20 predefined command groups. These were chosen for controlling the most frequent software programs or hardware devices, following the demands of the users. These applications include mainly:

- general access to any programs and files in the PC
- text input
- access to Internet services (e-mail and web browsing)
- control of a hardware devices mounted in the PC

The basic set of commands (namely the spelling words, names of keyboard keys, etc.) are prepared both for people with correct pronunciation as well as for those who have pronunciation problems. For example, letter A has a short pronunciation variant “A” and a long one “Alpha”.

3.2. Configuration of MyVoice commands and functions

In a voice control system for handicapped people, the users must have a chance to create their own commands and assign to them arbitrary sequences of actions. They should also have a possibility to change the pronunciation of individual commands arbitrary, because motor-handicapped persons often have various pronunciation deviations too. Due to this reason, all MyVoice commands, lexicons, actions and settings can be modified to a very large extent using the configuration tool. This can be accessed off-line as well as on-line (during the session) and it can be used for:

- adding, renaming or removing command groups
- modifying items in each command group
- generating phonetic transcription for voice commands, for most words (namely Czech-origin) this is done automatically, for others, at least an approximate pronunciation has to be specified
- assigning actions to each command, these can be composed of keystrokes, mouse moves and clicks, program launch or group switch

The MyVoice software package also includes tools for speaker adaptation and program configuration (e.g. selection of acoustic models). These are used in MyDictate too and are shared if both programs are installed on the same PC. In this case, it is possible to use a simple voice command to terminate MyVoice and launch MyDictate.

4. MyDictate - design and functions

The second system, MyDictate, shares several characteristics with MyVoice from the user perspective. For example, the main program's window (see Fig. 3) is always on the top of the screen. On the contrary, its appearance is different. Here, the main area is occupied by a list of next candidates from recognition. The size of this window can be changed from the contracted form (shown in Fig. 3) to a full-size form in which a complete list of all 262 commands is displayed.

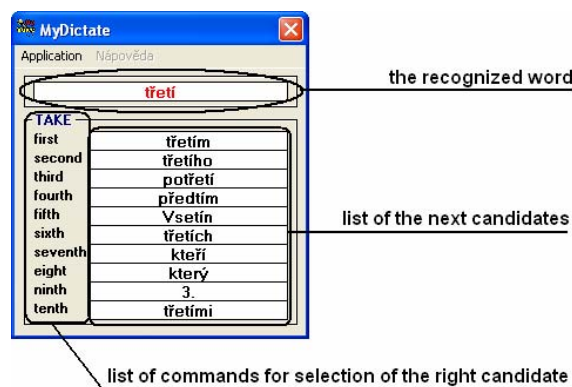


Figure 3: MyDictate's main window (in the contracted form)

4.1. Voice commands in MyDictate

A big proportion of MyDictate commands are similar to the ones of MyVoice. However, the main difference is that in MyDictate, their pronunciation form is longer than strictly needed, to improve their recognition accuracy. It is because all commands in MyDictate share the same vocabulary together with a half million of words. So for example, the command “left” in MyVoice corresponds to “cursor_left” in MyDictate. Briefly described, commands in MyDictate are prepared for:

- activation and deactivation of the output to the OS
- lexicon's modification and configuration of MyDictate
- control of cursor's movement, editing and selection of the text and spelling of words
- termination of MyDictate and launch of MyVoice (if these both systems are installed on the same PC)
- changing the size of the first character belonging to the word or phrase that was dictated as the last one
- selection of one candidate from the list of all next candidates – when this command is used then the word that was printed on the screen as the last one is replaced with the selected candidate
- deleting a word, two words or the last character

As mentioned above, MyDictate has several important functionalities (e.g. deleting the final characters and selection of next candidates) that allow correcting (respective replacing) each misrecognized word easily. For an inflective language like Czech, this is a very important feature because in it, a lot of acoustically closed words (with the same basic form) differ only in several letters at the beginning or end. Moreover, these commands also allow that the out of vocabulary (OOV) words can be dictated easily: the user only utters their basic form at first and then he/she corrects it using commands for deleting of characters and/or spelling.

4.2. Lexicon modifications

In MyDictate, the possibility of adding new words and modifying their pronunciation is critical in the same way as in MyVoice. The tool that we developed for this purpose can be again accessed off-line as well as on-line. After it is activated, a word or a phrase that was selected in the text beforehand is prepared for being included into the vocabulary automatically. As in MyVoice, a phonetic transcription is created for most

words by the system; however, the user also has the possibility to modify it. He/she can set the occurrence frequency of each new word (in terms of low, middle or high) to estimate its unigram factor too. The main advantage of our tool for lexicon modification is that all previously mentioned actions can be performed by single voice commands so that also motor handicapped users are able to modify the lexicon easily.

Moreover, during the phase of lexicon's modification, the vocabulary used for speech recognition is switched to a reduced variant containing only items (e.g. commands for control of cursor's movement or spelling words) that are necessary to create the phonetic transcription of the new word or edit its form for example. The advantage of this approach is that their recognition accuracy is thereby improved.

5. Performance evaluation

5.1. MyVoice

For evaluation of MyVoice, we used 1700 recordings obtained from the first user of MyVoice, a 16-year old quadriplegic girl. During this experiment, words from all groups were put together to one vocabulary containing 539 items.

At first, a gender dependent (GD) model was applied. The resulting low recognition accuracy (83 %), was caused by the fact that the girl had a speech defect which was connected with her physical handicap. For persons with standard pronunciation, the baseline recognition rate with GD models was above 97 %. Then we used our speaker adaptation (SA) module based on the combination of MAP [3] and MLLR [4] and prepared set of 320 adaptation words. These were the most important voice commands, most frequent Czech words and were selected to cover all Czech phonemes. After that, the recognition rate increased to 93 %. This proved the ability of the system and its SA module to perform satisfactorily even for a speaker with non-standard pronunciation.

5.2. MyDictate

During the evaluation process of MyDictate, we focused not only on system's accuracy, but also on other aspects that are connected with its practical usability (like the speed of dictation). For adaptation, we used again a set of 320 adaptation words, which was formed using similar guidelines as for MyVoice.

In the first experiment (Tab. 1), two non-handicapped persons were asked to dictate various newspaper articles.

Table 1: *Practical results of text dictation with MyDictate*

type of the article	user A		user B	
	news	sport	news	sport
# of words (without punctuation)	280	258	362	261
total # of pronounced words (including punctuation)	491	388	506	406
# of control commands used	83	55	57	49
# of OOV words	9	3	1	8
speed of dictation [chars per minute]	129	137	156	127
GD/SA model's accuracy [%]	88/91	90/92	94/96	92/93

Results of this experiment showed that although the average recognition accuracy of the adapted systems is over 92 %, the number of words that must be pronounced to dictate a given text with correct punctuation is much higher – about 50 % more in average – than the real number of words in the text (not including punctuation as standard). This problem happens mainly for texts containing a high number of OOV words – see Tab. 1. In this case, the user usually tries to say the word at first and then he/she has to delete or correct the word form that was misrecognized. Finally, the user has to employ several more voice commands, when he/she wants add the word to the vocabulary. All these facts decrease the speed of dictation on the level of 140 chars per minute, which can be interesting and useful namely for handicapped persons but is low in comparison with skills of a professional typists.

Finally we present a preliminary experiment that shows My Dictate's performance for a 18-year old paralyzed boy. His paralysis was caused by an abnormally strong muscle tension in his body which was projected onto his vocal chords and affects his pronunciation. Due to this reason, his baseline recognition accuracy with GD models was only about 83 %. After adaptation, it improved on level of 88 % which is closer to the accuracy that was reached by persons without any speech defect (Tab. 1).

6. Conclusions

In this paper, we presented a general speech recognition engine that was created to develop two voice controlled aids for Czech handicapped users: MyVoice for voice control of PC and MyDictate for text dictation. These give Czech handicapped people not only an access to a PC and its programs or an opportunity to communicate with other people through internet, but also a chance to find a job aimed on work with computers. It is the reason why MyVoice (developed in 2005) is already used by several tens of handicapped people and why we were asked (early in 2007) to develop the complementary system MyDictate. In the future, we plan to extend functions of both these systems according to the most frequent requirements of their target users (e.g. voice feedback for persons with visual impairment).

7. Acknowledgements

The research was supported by the Grant Agency of the Czech Academy of Sciences (grant 1QS108040569).

8. References

- [1] Nouza, J., Nouza, T., Červa, P., "A Multi-Functional Voice-Control Aid for Disabled Persons", Proc. of Specom 2005, pp. 715-718, Patras, Greece.
- [2] Nouza, J., Psutka, J., Uhlíř, J., "Phonetic Alphabet for Speech Recognition of Czech", Radioengineering, vol.6, no.4, pp.16-20, Dec.1997.
- [3] Gauvain, J.L., Lee, C.H., "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. SAP, Vol. 2, pp. 291-298, 1994.
- [4] Gales M.J.F., Woodland P.C., "Mean and Variance Adaptation Within the MLLR Framework", Computer Speech & Language, Vol. 10, pp. 249-264, 1996.