speech analyse, voice control,
inefficiency persons,
MSAA technology

Piotr PORWIK*

USER VOICE IDENTIFICATION IN COMPUTER APPLICATIONS

Producers of computer devices very often fit their products with functions, which can be used by persons with some inefficiency. Such conveniences can be applied for blind or partially sighted users or for persons with limb paresis. The application should be user-friendly with wide spectrum of adjusts. Computer users have wide choice of software speech analysers and synthesisers that may help them during working on computer. In this paper Microsoft Windows application with firmware MSAA technology is proposed. In such program all application events by means of user's voice are controlled. Practical, investigations of voice recognition have also been presented.

1. INTRODUCTION

The operating system is the important part of each computer system. The last may be divided into the following parts: hardware, operating system and software. The hardware (processor, memory and I/O devices) makes up main resources of a computer system. The software (compilers, database systems, etc.) characterises the ways of usage of these resources to solve tasks made by users. In many cases, software producers take into consideration specific needs of some users group – with many dysfunctions. For this reason Microsoft Windows operating system has options, which mainly are used by persons with disabilities: magnifier function or built-in text-to-speech module. In particular, can be applied functions of "Microsoft Active Accessibility" (MSAA) that standardise the procedures of interpretation of individual objects on the screen. It is new technique, which allows to control application by reading active fields, which are actually presented on the screen. Because application objects' by programmer can be controlled – it is possible to connect it with hardware/software speech analysers. In such case, utterance of isolated words of user can be recognized and used in steering of user's application. Hence, application management by means of user's voice can be done. In this paper, all considerations concern of applications where Polish language is used and appropriate isolated words are recognized. Proposed in the paper software for Polish user's is dedicated, hence in all discussed examples Polish version of operating system is performed and also all reporting operating messages in Polish language are shown. It can be noticed that with no major obstacles described application control method can be adopted to any another language.

* Institute of Informatics, Silesian University, Będzińska 39, 41-200 Sosnowiec, Poland

2. MICROSOFT ACTIVE ACCESIBILITY (MSAA)

In modern computer applications, where graphical user interfaces (GUIs) are used, object-oriented programming is applied. GUI is now firmly established as the preferred user interface for end users. In such applications, graphical user's window can include: buttons, radio-buttons, check boxes, combo-boxes, icons, pull-down menus, text fields, etc. In order to convey meaningful information from user interface, we must be able to access that information from the application. Solution to this problem is Microsoft Active Accessibility (MSAA), which has been available as an add-on since Windows 98. MSAA is a technology that provides a standard, consistent mechanism for exchanging information between applications and assistive technologies. For example, MSAA allows applications to expose screen readers to the type, name, location, and current state of all objects and notifies screen readers of any Windows event that leads to a user interface change. Although it is not the only way to communicate with assistive technology, MSAA allows programmers to support a broader variety of applications without custom programming for each one. The number of applications that support MSAA is growing, although there still are many popular applications that do not support it.

The important feature of the MSAA tool is possibility to use and controlling these features in special applications. A set of accessibility features of application objects make it easier for persons with disabilities to use computers. The MSAA is a set of interfaces and APIs that provides a reliable way to expose and collect information about Microsoft Windows-based user interface (UI) elements. Using this information, programmers and users can represent the UI in alternative formats, such as speech, Braille or voice command and control applications can remotely manipulate the interface. The MSAA technology can be used in conjunction only with Windows-based controls.

The main idea of the MSAA interface base on special functions designed for UI elements. If UI elements are accessible for MSAA interface – are redirected to *IAccessible* interface and *child* ID. It is enough to UI object description. The simple window (it is part of global application) and accessible MSAA features – as description of the window elements presents Fig.1a.

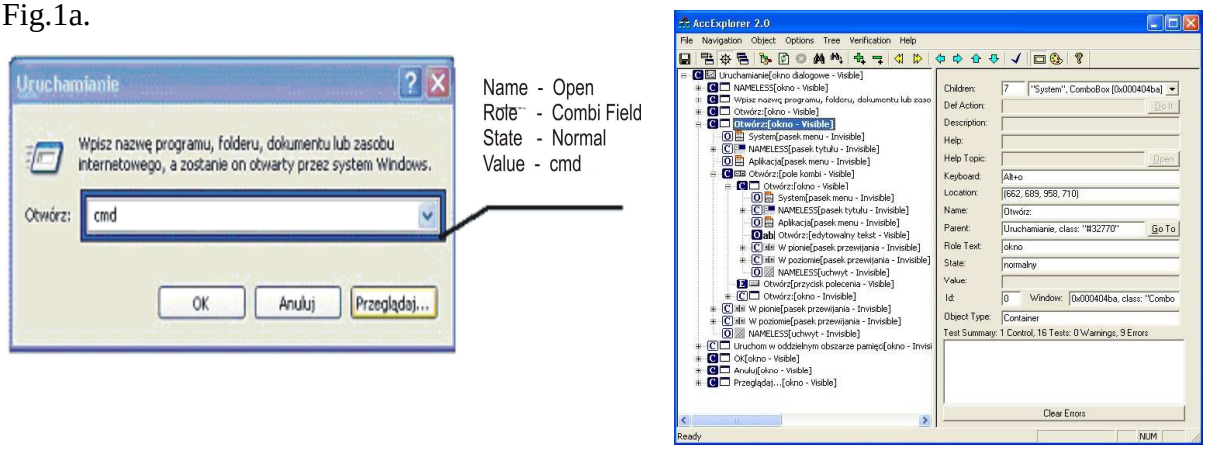

Fig.1 An example of window, for which accesible MSAA features can be analyzed and recognized by *IAccessibe* interface (a) and MSAA assistant tool: the AccExplorer program (b)

Unfortunately, designing of the MSAA applications is difficult. Hence, appropriate tools are freely available as auxiliary applications. One from such applications is AccExplorer program. The AccExplorer allows to indicate all features which are accessible for

programmer. For any object, which can be modified in user's application, appropriate MSAA properties are shown as tree of inherited features (Fig. 1b). If application object by means of the cursor will be indicated, then the AccExplorer window automatically is visible on the screen and associated with MSAA feature. For all MSAA features user can propose appropriate model of word, then this one is recorded and stored in database. Exhaustive considerations about MSAA technology can be found in [5].

3. HUMAN VOICE AS COMPUTER CONTROL TOOL

There are two uses for speech recognition systems:
Dictation – translation of the spoken word into written text, and computer control – control of the computer and software applications by speaking commands.
Speech recognition is one of the desired assistive technology systems. After recognition, such control is a natural and easy method of accessing the computer. From this reason computer control by means of human voice is very attractive for many user's. Speech-recognition programs do not understand what words mean, but isolated words can be recognized and used as context tool. This information helps the computer choose the most likely word from database. Speech-recognition software, on the other hand, works best when the computer has a chance to adjust to each new speaker. The process of teaching the computer to recognize voice is called training. Because voice commands have to be unambiguous, therefore different recognition methods should be simultaneously used. Extraction of words from speaking commands is a very difficult task; therefore in proposed approach different methods of speech recognition have been implemented. The procedures for features extraction are as follows:
Linear Prediction Coefficients (LPC):

- Autocorrelation Coefficients (AC),
- Reflection Coefficients (REF),
- Linear Prediction-based Cepstral Coefficients (LPCC).

Mel-Frequency Cepstral Coefficients (MFCC).
Hidden Markov Models (HMM).
The diagram of recognition of isolated words presents Fig. 2, where connections between blocks are shown. In the first stage appropriate isolated words are expressed, registered and its parameters are computed. Each recognised word, in database directory is stored.

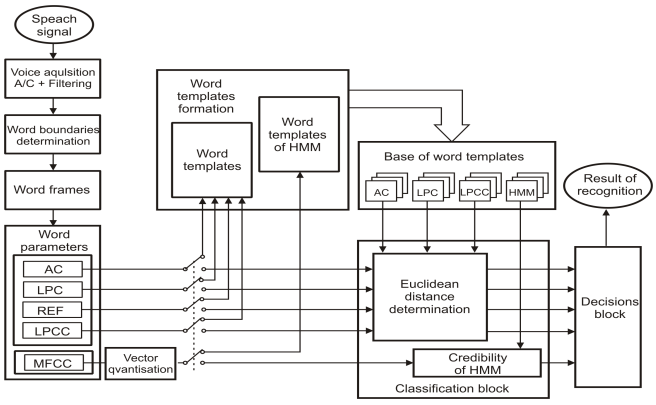

Fig. 2. Block scheme of words recognition system

The next section presents the experimental investigations conducted into the relative effectiveness for speaker recognition by means of the above mentioned feature types.

4. AN EXTRACTION OF ISOLATED WORDS FROM THE SPEECH SEQUENCE

The classical approach to speaker recognition is through the use of short-term spectral templates. The approach involves applying an appropriate analysis to a spoken utterance to generate a sequence of short-term spectral feature vectors. Such templates have been found to contain significant voice characteristics, and may therefore be used to effectively discriminate amongst speakers. Amongst various types of speech features, LPCC and MFCC have been found to be superior for speaker recognition [1,2,3,4]. Voice acquisition by embedded sound card and microphone has been registered. Acquisition parameters can be individually selected, depending on sound cart quality. In presented experiment 8 kHz frequency sampling has been established, each sample in 8 bit resolution mode was registered. Time of the sample recording was 3*s*. Each sample was three times registered, hence $3\times 8000 = 24000$ samples per word in database have been stored.

PRELIMINARY VOICE FILTERING (pre-emphasis).

Registered by microphone speech signal is filtered. In this process, non recursive Finite Impulse Response (FIR) filter has been used. Operation of the FIR filter can be described by equation:

$$\overline{y}(n) = y(n) - a \cdot y(n-1), \tag{1}$$

where:

n – number of current sample,
$\overline{y}(n)$ – sample of digital signal after FIR filtration,
$y(n)$ – sample of digital signal before FIR filtration,
a – filtration coefficient, $a = 0{,}937$.

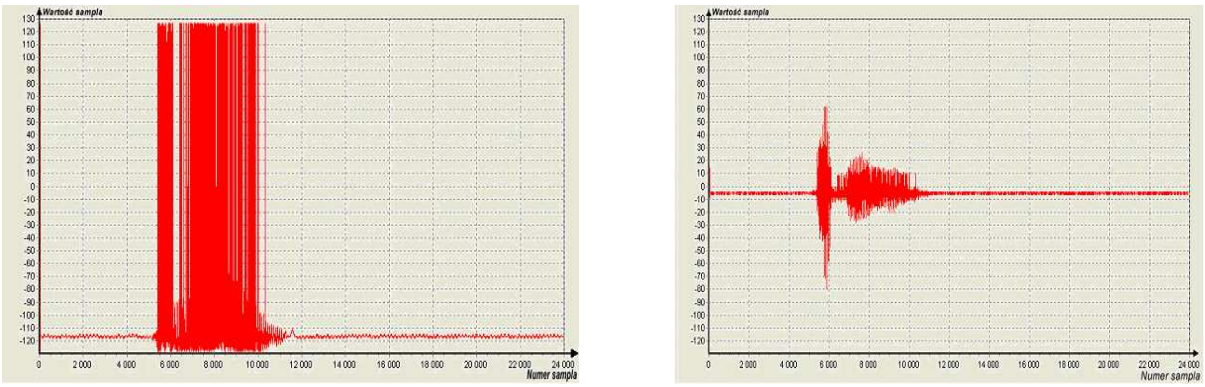

Fig. 3 The word "Anuluj" (a) and this word after FIR filtering (b)

From fig. 3 can be observed that after FIR filtering, signal dynamics is reduced, low frequencies are attenuated and high frequencies occur in speech signal are peaked.

FRAME ENERGY AND NORMALISATION

Speech signal (Fig. 4b) into frames is divided and each frame has $N = 300$ signal samples. Such number of samples is equal of 37ms of speech recording. For each frame, energy of signal is computed:

$$E(l) = \log\left(\sum_{n=1}^{N} \overline{y}_l^2(n)\right),\ 1 \le l \le K \tag{2}$$

and energy normalisation is performed:

$$E_{Norm}(l) = \frac{E(l)}{\max\{E(1),...,E(K)\}} \tag{3}$$

where:

l – number of current frame,

K – number of all frames,

$E(l)$ – energy of l^{th} frame,

$E_{Norm}(l)$ – normalised energy of l^{th} frame.

Fig. 5a presents an energy distribution of the word "Anuluj" and a normalised distribution of this same word (Fig. 5b). Additionally, in Fig.5 two horizontal lines are shown because after normalisation, the two (upper and lower) energy levels are established. The upper level has value 0,6 and determine beginning of the word. The lower level has value 0,5 and determine end of word. The energy levels were experimentally selected. On the basis of energy levels, boundaries of the word can be fixed. Range of the word allows to determine silent and noise places, which occur during utterance.

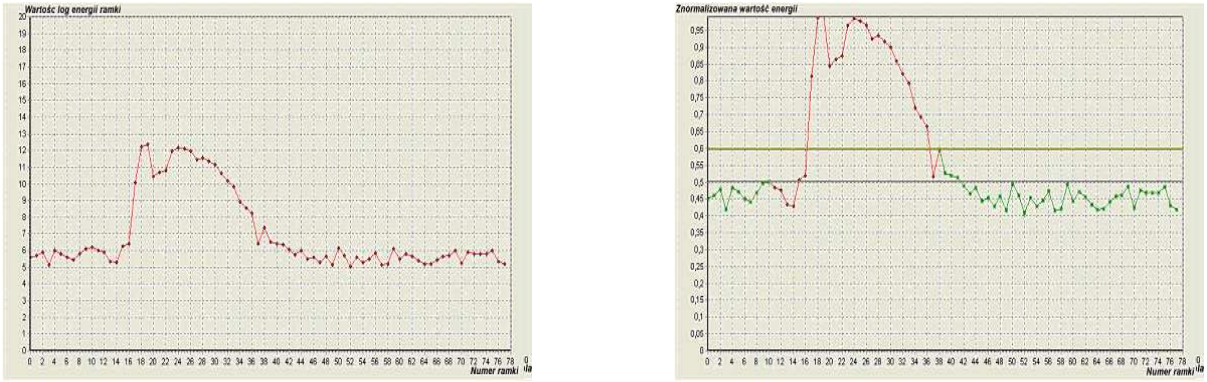

Fig. 5 Energy of the word Anuluj" with log scale (a) and distribution of normalised energy of this word (b)

In this way, an isolated word (without a noise and silence) is determined (Fig. 6a). In this figure by means of the green colour range of the word has been shown. The red colour indicates noise or silence level. In the next stage, isolated word is again divided into frames (Fig. 6b). Many authors report, that range of the time frame should be 20–40*ms* [2,3,5]. In proposed experiment time of frame is 32*ms*, what is equal to N samples of signal. Typically it is an integer power-of-2, such as $N = 2^8 = 256$. N represents the width, in samples, of a discrete-time window function. Such signal can be treated as quasi-stationary. In the next stage, for each frame the Hamming window is applied. Windowing operations are chosen to improve of quality of a signal. When using FFT for spectral analysis a sample belongs to one analysis window. When using windowing, samples at the boundaries are attenuated.

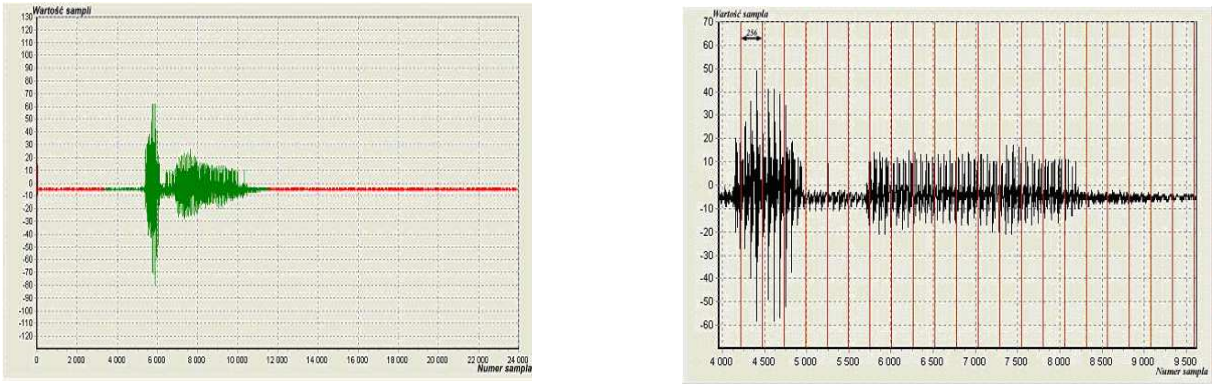

Fig. 6 Isolated word (green colour) (a) and the same re-scaled word divided into *N* frames (b)

To reduce the effect that these samples become less important for the result, normally windows are overlapped. So samples between two blocks are attenuated, but they belong to two blocks: their influence is still (nearly) the same as samples which are not attenuated. Windowing for frame signal $\bar{y}(n)$ can be described as:

$$y^*(n) = \bar{y}(n) \cdot w(n), \quad w(n) = 0{,}54 - 0{,}46 \cdot \cos\left(2\pi n/(N-1)\right), \qquad 0 \le n \le N \tag{4}$$

where: $w(n)$ is the Hamming window.

5. USER VOICE PARAMETERS

In proposed in this paper approach different voice features are computed (Fig. 3) on the basis of frame of word (Fig.6b). The procedures for feature extraction are as follows [1,3]:
LPC – Linear prediction coefficients are obtained for each frame using the Levinson-Durbin recursive method [2,4]. These coefficients are then converted to cepstral coefficients.
MFCC – Fast Fourier Transform is computed for each frame and then weighted by a Mel-scaled filter bank [4]. The filter bank outputs are then converted to cepstral parameters by applying the inverse discrete Fourier transformation. As has above been stated, the speech signal is sampled at a frequency of 8 KHz and is processed in the LPC algorithm in blocks of 2560 samples. Therefore, the LPC algorithm must extracted the required parameters characterizing the 256 samples of speech, each 32*ms*. Samples are high pass filtered and windowed using a Hamming window. The obtained data samples are used as the input to the autocorrelation detection blocks. The Levinson-Durbin algorithm is then used to solved a set of p linear equations. Reflection coefficients (REF) by the Levinson-Durbin algorithm are also determined. These linear equations are functions of the sequence autocorrelation and the solution is the set of coefficients $a(i)$. From values $a(i)$, cepstral coefficients can be calculated:

$$c(1) = a(1), \quad c(k) = a(k) + \sum_{m=1}^{k-1} \frac{m}{k} \cdot c(m) \cdot a(k-m), \quad 2 \le k \le p$$

$$c(k) = \sum_{m=1}^{k-1} \frac{m}{k} \cdot c(m) \cdot a(k-m), \quad k > q \tag{5}$$

In the last stage, MFCC parameters were computed.:

$$c_{mel}(k) = \sum_{n=1}^{L} \ln(\tilde{S}(n)) \cdot \cos\left(\frac{\pi k}{L}(n-0{,}5)\right), \; k = 1,2,...,q \tag{6}$$

where:

q – number of cepstral coefficients,
L – number of mel filters,
$\tilde{S}(n)$ – estimation of the FFT power spectrum,

Mentioned parameters (coefficients) constitute the vector cepstral coefficients of the isolated word. For each vector, Euclidean distance between recognized word and pattern from database is calculated. Speech recognition can be treated as stochastic process and can be described with the aid of hidden Markov modelling technique (HMM). Speech is a continuous stream of acoustic information. Even if we assume that the talker must stop sometimes, the possible utterances vary in length and their number is practically unlimited. A possible solution is to trace the problem by the HMM technique. In this tecnique for a given output

sequence, the most likely set of state transition and output probabilities is searched. It can be achieved by the Baum-Welch (B-W) algorithm. If isolated word from speech sequence is determined, then on the basis of the B-W algorithm, and HMM word models, stored in database, highest probability of apperance of such word is calculated. Finally, if word is recogited, appriopriate MSAA procedure is activated.

6. INVESTIGATION RESULTS

In carried out experiment was investigated whether text utterance has actually been produced by the speaker associated with the best-matched models from database, or by unknown speaker outside the registered set of features.

Experimental set up

Method	**Parameter**
Nature of speech data	Isolated words
Features under consideration	All LPC parameters, MFCC, HMM
Speaker modelling	Spiker_Demo-2.6 – polish male voice
Number of registered speakers	1
Number of unknown speakers	1
Number of known utterances	500
Training data duration	3 times specific word per 3 seconds
Performance measure	Identification error

Quality of identification by means of the next equation was checked:

$$IDE(ntification) = \{1 - [(correct\ identifications) / (total\ tests)]\} \times 100\% .$$

Type of speaker	**IDE** (%)
Registered speaker	0,04
Unknown speaker	0,75

On the basis of recognised words (speakers) voice control is performed, and MSAA procedures allow in practice carried out such control. From investigations follow that recognition level is satisfactory on condition that short training procedure will be used.

BIBLIOGRAPHY

[1] Deller J.R. et al. *Discrete-Time Processing of Speech Signals*, Macmillan Pub. Company, 2000.
[2] Rabiner L. R., Juang B.H., *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
[3] Huang X.D.,AceroA., Hon H-W., *Spoken language processing*. New York, Prentice Hall, 2001.
[4] Zieliński T.P., *Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań*. WKŁ, 2005.
[5] *Microsoft Active Accessibility*. Version 2.0. On line Microsoft documentation.