UDK 004.522:681.516.33

EMBEDDED SYSTEM AND SPEECH RECOGNITION

Bučko R., research, Ing., Assist. Prof. (Eng.), D. Kováč, Prof. (Eng.)
Technical University of Košice
Letná 9, 042 00 Košice, Slovak Republic
E-mail: radoslav.bucko@tuke.sk, dobroslav.kovac@tuke.sk
I. Konokh
Kremenchuk Mykhailo Ostrohradskyi National University
vul. Pershotravneva, 20, 39600, Kremenchuk, Ukraine
E-mail:pks@kdu.edu.ua

This paper describes an embedded systems and signal processing for recognition of voice commands for the car's equipment control such as GPS, radio, air-conditioning etc. or robotic system.

Key words: embedded system, signal processing, Euclidean distance, spectrogram.

Introduction. In present time the recognition of spoken speech is highly developed. Communication using verbal speech is the most basic and natural form of information transfer between people. With new communication and information technology it is becoming necessary to use verbal speech for communication with computer.
Most research is focused on using the English language for such communication, however our research is aimed for using the Slovak language.

Research of recognition of spoken speech on our department is oriented on recognition of simple instructions by spectrogram. These instructions are used for car's equipment control such as GPS, radio, air-conditioning or robotic system with embedded systems.

Problem statement. Some problems with recognition of spoken speech are:

a) speaker's voice can be different in various conditions,
b) different speakers have different voices,
c) changing environment causes trouble for speech,
d) recorded voice can be degraded by quality of microphone or by distance from it. [1]

An embedded system is a computer system designed to perform one or more dedicated functions often with real-time computing constraints. One or more main processing cores control embedded systems. They are typically represented either by microcontrollers or digital signal processors (DSP). The program instructions written for embedded systems run with limited computer hardware resources: little memory and operating output. Because of this, it is important to optimize the acoustic signal processing used by embedded systems for speech recognition. [2]

Mechatronic system Teach-Robot.

For control with simple instruction are selected Teach-Robot. Mechatronic system Teach-Robot is angular arm with 5 axles and 6 DC-motors (Table 1). Teach-Box provides manual control of Teach-Robot and provides communication between Teach-Robot and personal computer (Fig.1). [3]

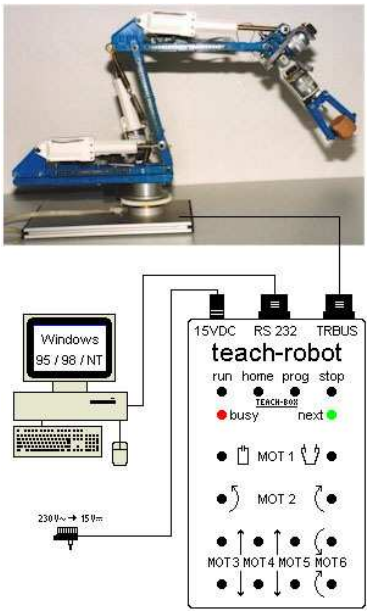

Figure 1 – Teach-Robot

Table 1 – Moving specifications

Meaning	Motor	Number of pulses	Angle
grip of gripper	M1	70	60°
rotation of gripper	M2	130	200°
wrist up/down	M3	420	90°
rotation of upper arm	M4	420	90°
rotation of lower arm	M5	350	80°
rotation of body	M6	700	320°

Signal processing

Speech sounds are created by vibratory activity in the human vocal tract. Speech is normally transmitted to a listener's ears or to a microphone through the air, where speech and other sounds take on the form of radiating waves of variation in air pressure around an average resting value at sea level of about 100,000 Pa [4].

The basic principle of most methods for acoustic signal processing is the assumption that its properties are changing slowly. Methods called short-term analysis separated and processed segments of speech

signal. These segment are micro segments which are represented by the time segment of 10 to 30 ms (our system 10 ms and overlap 5 ms) . Because these micro segments are connected or can overlap each other we will get the sequence of numbers, which describes the speech.

Speech signal is recorded mostly by microphone, so the analogue signal is recorded. Analogue cycles are digitalized, that the continuous signal is represented by sequence of numbers. This process is called pulse code modulation.

Pulse code modulation consist of two operations:

- sampling in time,
- quantization.

Sampling in time – samples are taken from continuous signal are in periodic moments $t_n=n.T$ which sizes corresponds to immediate values of continuous signal in sampling time t_n. T is the sampling period and $n=0,1,...,\infty$. [3]

According to Shannon´s sampling theorem the frequency of sampling f_v must be twice as the maximum frequency of analogue signal f_m:

$$f_v \geq 2f_m \tag{1}$$

Quantization is the operation, which allows the change of signal with continuous variable to signal with finite number of values.

Processing *by time*

Most methods of short term analysis in time can be described by following equation:

$$Q_n = \sum_{k=-\infty}^{\infty} \tau(s(k)) w(n-k), \tag{2}$$

where Q_n is the short time characteristics, $s(k)$ is the sample of acoustic signal get by pulse code modulation in time k, $\tau(s(k))$ is the transformation function a $w(n)$ is the weight sequence (or window) which chose the samples $s(k)$.

Hamming windows

Hamming´s windows are used when processing in time. Hamming´s window is defined as (fig.2):

- $w(n)=0,54-0,46\cos[2\pi n/(N-1)]$ for $0 \leq n \leq N-1$,
- $w(n)=0$ for other n.

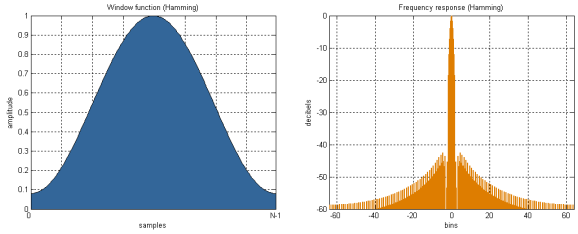

Figure 2 – Hamming window

Hann windows

Hann´s window is defined as (fig.3):

- $w(n)=0,5\{1-\cos[2\pi n/(N-1)]\}$ for $0 \leq n \leq N-1$,
- $w(n)=0$ for other n.

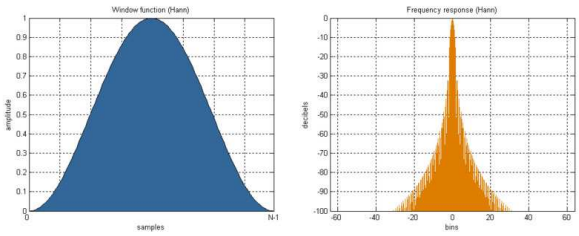

Figure 3 – Hann window

Rectangular window

Rectangular window (sometimes known as Dirichlet window) is defined as (fig.4):

- $w(n)=1$ for $0 \leq n \leq N-1$,
- $w(n)=0$ for other n.

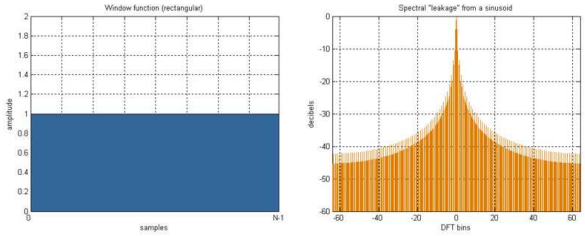

Figure 4 – Rectangular window

Speech detection

Big problem with speech recognition is how to determine when the instruction is spoken. Two basic short-time analysis functions useful for speech signals are the short-time energy and the short-time zero-crossing rate. These functions are simple to compute, and they are useful for estimating properties of the excitation function in the model.

Function of short-time energy can be described by the following equation:

$$E_n = \sum_{k=-\infty}^{\infty} [s(k) w(n-k)]^2, \tag{3}$$

where $s(k)$ *is* the sample of acoustic signal get by pulse code modulation in time k and $w(n)$ is in our system windows (Hann, Hamming or Rectangular) with length of micro segment 10 ms and sampling rate 8 kHz.

The short-time zero crossing rate is defined as the weighted average of the number of times the speech signal changes sign within the time window. Function of zero crossing rate can be described by the following equation:

$$Z_n = \sum_{k=-\infty}^{\infty} |sgn[s(k)] - sgn[s(k-1)]| w(n-k), \tag{4}$$

, where

- $sgn[s(k)]=1$ for $s(k) \geq 0$,
- $sgn[s(k)]=-1$ pre $s(k)<0$,
- $w(n)$ represents some windows.

Euclidean distance

The simplest and most efficient method for comparison of 2 vectors is the calculation of their Euclidean distance:

$$d(a,b) = \sqrt{\sum_{k=1}^{n} [a_k - b_k]^2} \tag{5}$$

where $A=[a_1,a_2,...,a_n]$,
$B=[b_1,b_2,...,b_n]$.
The size of both vectors has to be the same.

Spectrogram

We have chosen spectrogram for our method of speech recognition with embedded systems.

A spectrogram is a time-varying spectral representation (forming an image) that shows how the spectral density of a signal varies with time.

The vertical axis shows positive time towards the up, the horizontal axis represents frequencies, and the colors represent the most important acoustic peaks for a given time frame, with red representing the highest energies. [8]

A spectrogram is calculated from the time signal using the short-time Fourier transform (STFT). In our research we have chosen three windows for STFT – Hann (Fig.5), Rectangular (Fig.6) and Hamming window (Fig.7). This spectrogram with different windows is similar, but we can see differences that we can use for improvement of our system.

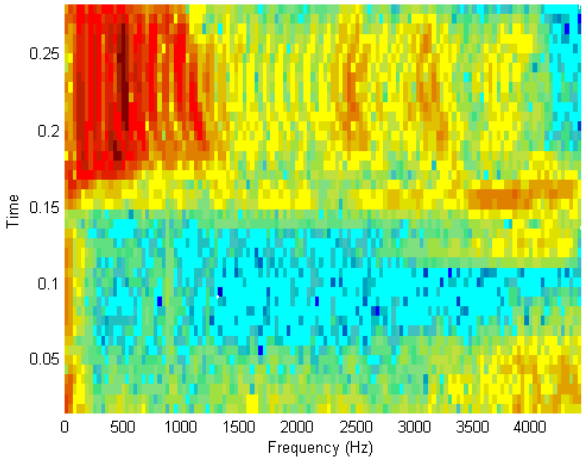

Figure 5 – Spectrogram (Hann window) of word “stop” spoken by first person

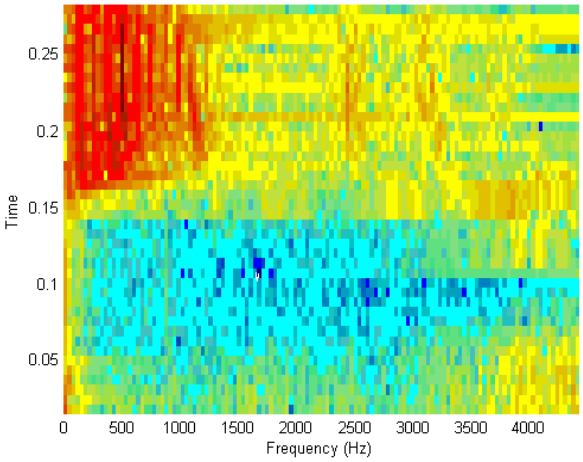

Figure 6 – Spectrogram (rectangular window) of word “stop” spoken by first person

Figures (Fig.7, Fig.8) are showing the same word (word “stop”) spoken by two people. We can see similarities there. Fig.9 is showing different word (word "vpravo"). We can see that the spectrographic picture is clearly different, which is very important for our recognition system.

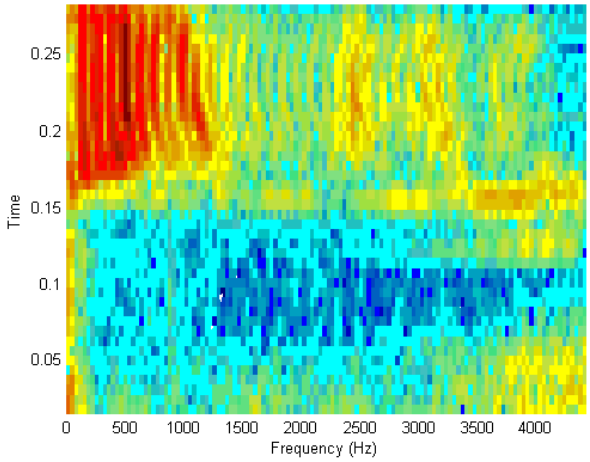

Figure 7 – Spectrogram (Hamming window) of word “stop” spoken by first person

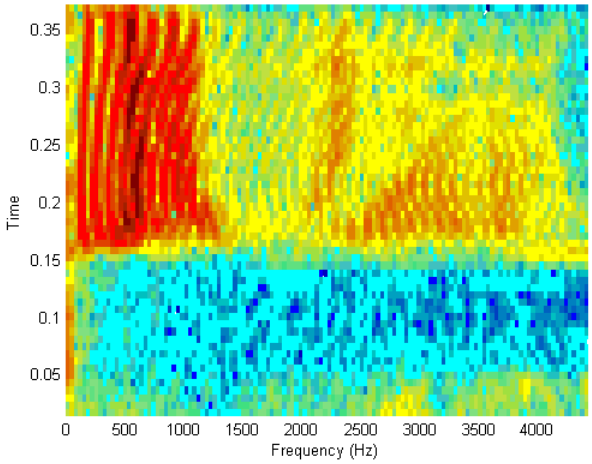

Figure 8 – Spectrogram of word “stop” spoken by second person

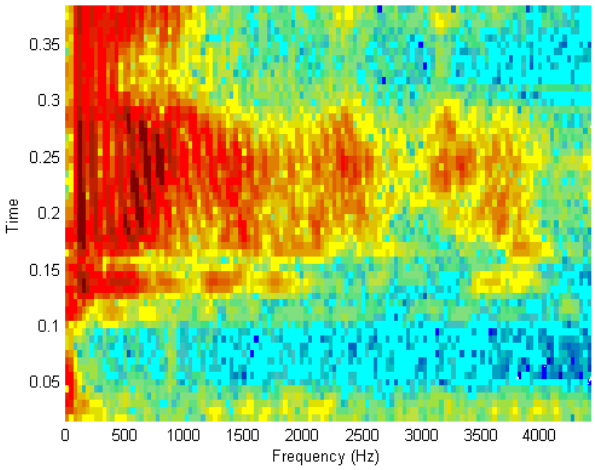

Figure 9 – Spectrogram of word “vpravo” spoken by second person

We have divided the spectrogram to several sectors, in order to ease the computing process. The final value is made by arithmetic mean of these sectors (Fig.10).

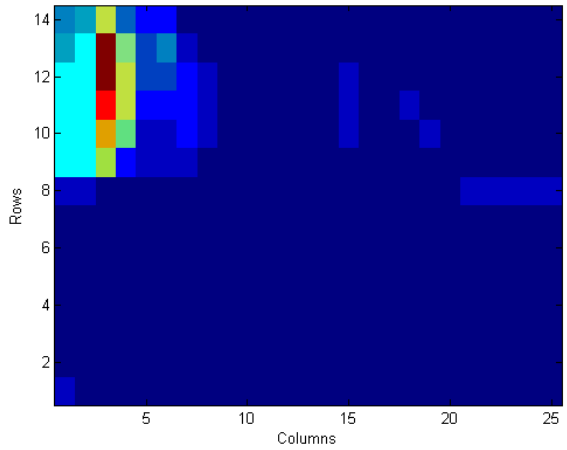

Figure 10 – Spectrogram of word “stop” with arithmetic mean of sectors spoken by first person

Normalization of spectrogram

Figures (fig.10, fig.11) show the different word (word “stop” and word “vpravo”) spoken by one person. We can see that number of rows is unequal-14 rows in spectrogram of word “stop” and 19 rows in spectrogram of word “vpravo” . If we want to calculate the Euclidean distance of these two spectrogram (word “stop” and word “vpravo”) we must normalize one of them to the same number of rows. We can normalize the higher amount of rows (19) to lowe amount (14) using the crop or we can fill the missing rows to get the higher amount of rows, when needed. We have chosen the second method, where we filled the rows 15-19 with blank rows (fig.12)

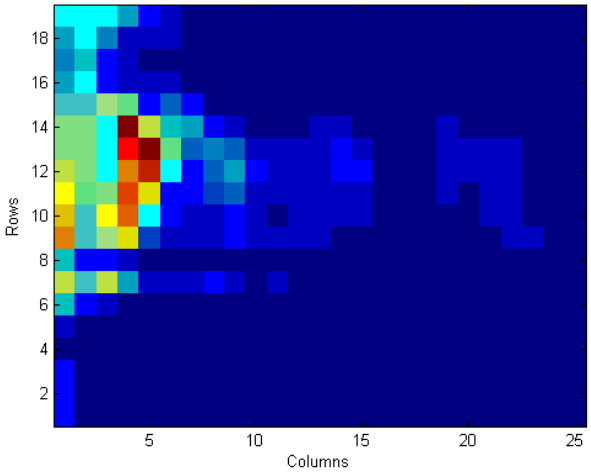

Figure 11 – Spectrogram of word “vpravo” with arithmetic mean of sectors spoken by first person

Euclidean distance of spectrogram of words „vpravo“ and „stop“ is 10,7857. (fig.14) This distance is higher as Euclidean distance of spectrogram of two same words „stop“ = 4,27785.

For improved recognition of commands we need to use 2 things:

1) calculated Euclidean distance from spectrograms, that are calculated using STFT with 3 windows – Hann, Hamming and Rectangular windows.
2) second arithmetic mean.

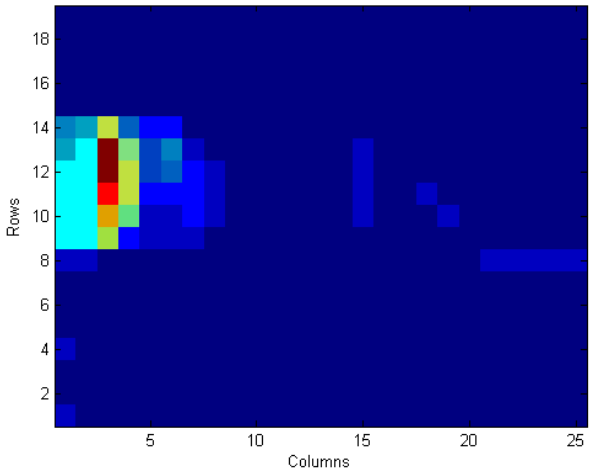

Figure 12 – Spectrogram of word “stop” with arithmetic mean of sectors spoken by first person after normalization

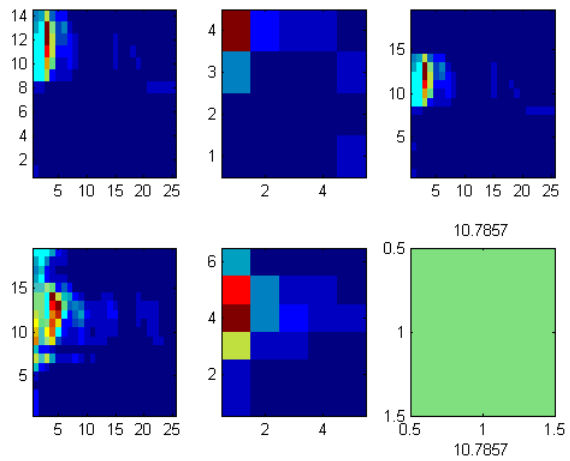

Figure 13 – Comparison of spectrogram of word “stop” - up and word “vpravo” - down with arithmetic mean, second arithmetic mean, normalization of spectrogram and Euclidean distance

Conclusions.

Spectrograms together with Euclidean distance offer interesting way of speech recognition with using only a limited hardware resource of embedded systems.

Acknowledgment.

The paper has been prepared by the support of Slovak grant project KEGA 003-003TUKE-4/2010.

We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centre of Excellence of the Integrated Research & Exploitation the Advanced Materials and Technologies in the Automotive Electronics ", ITMS 262201200.

REFERENCES

1. J. Psutka, “Comomunication with PC using spoken speech“ ACADEMIA, Prague, 1995.

2. Embedded system, http://en.wikipedia.org/wiki/ Embedded_system, 18.7.2010.

3. Čuchran, J.: “Digital transmission system“. STU, Bratislava, 2008.

4. Carmell, T.: “Spectrogram Reading”, http://cslu.cse.ogi.edu/tutordemos/SpectrogramReading/ spectrogram_reading.html, 7.4.2011

5. Kůs, V.: Influence of semiconductor converters on power system (in Czech). BEN technical literature, Prague 2002, 184 pages, ISBN 80-7300-062-8B.I.,

6. J. Psutka, “Speaking Czech with computer” ACADEMIA, Prague, 2006.

7. Homemade Speech Recognition with NETcf, http://www.mperfect.net/noreco/, 15.2.2010

8. Hagiwara R., How to read a spectrogram, http://home.cc.umanitoba.ca/~robh/howto.html, 7.4.2011

Стаття надійшла 06.06.2011 р.
Рекомендовано до друку к.т.н., доц.
Гладирем А.І.

ВСТРАИВАЕМАЯ СИСТЕМА И РАСПОЗНАВАНИЕ РЕЧИ

Р. Бачко, доц., Д. Ковач, к.т.н., проф.
Технический университет Кошице
ул. Летна, 9, 04200, Кошице, Словакия
E-mail: tibor.vince@tuke.sk , dobroslav.kovac@tuke.sk
И. С. Конох
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, 39600, Кременчуг, Украина
E-mail: pks@kdu.edu.ua

В статье описана обработка сигналов во встраиваемых системах для распознавания голосовых команд и управления такими системами автомобиля как gps, радио, кондиционер, а также роботизированными системами.

Ключевые слова. Встраиваемые системы, обработка сигналов, евклидово расстояние, спектрограмма.

ВБУДОВАНА СИСТЕМА І РОЗПІЗНАВАННЯ МОВИ

Р. Бачко, доц., Д. Ковач, к.т.н., проф.
Технічний університет Кошице
вул. Летна, 9, 04200, Кошице, Словакія
E-mail: tibor.vince@tuke.sk , dobroslav.kovac@tuke.sk
І. С. Конох
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, 39600, м. Кременчук, Україна
E-mail: pks@kdu.edu.ua

У статті описана обробка сигналів у вбудованих системах для розпізнавання голосових команд та управління такими системами автомобіля як gps, радіо, кондиціонер, а також роботизованими системами.

Ключові слова. Вбудовувані системи, обробка сигналів, евклідова відстань, спектрограма.