Neural network classification of gunshots using spectral characteristics

MILAN NAVRÁTIL, VOJTĚCH KŘESÁLEK, PETR DOSTÁLEK
Department of Electronics and Measurement
Tomas Bata University in Zlín, Faculty of Applied Informatics
nám. T.G.Masaryka 5555, 760 01 Zlín
CZECH REPUBLIC
{navratil; kresalek; dostalek}@fai.utb.cz http://www.fai.utb.cz

Abstract: - The paper describes neural network classification of specific audio sources into given categories. Audio sources are represented by various gunshots, from handguns or big bore guns. This article is a follow-up to an existing system for localization of audio sources from security and military areas. The question of successful classification lies in the convenient discrimination of the feature vector in the feature vector space. A set of feature vectors based on power spectral density is evaluated a tested for the best classification of gunshots.

Key-Words: - Neural network, classification, gunshots, spectral analysis, feature vector

1 Introduction

Nowadays, applications using artificial neural network algorithms are expanded a commonly utilized. It is due to huge technical development of the computers and computing system in general, especially last decades.

The need of classification of given type of weapon arises from a system for localization of audio events in military area using microphone array [1]. It is very useful to have information not only about direction and distance but also about specific weapon category which the analysed audio event could belong to. Moreover, this can help during investigation of crime incidents in common life where audio evidence is available.

In general, classification of audio data using machine learning techniques is well-studied problem. A number of methods have been proposed to classify various types of sounds, for example, music, speech, and others. It has been attempted before by researchers whose task was to distinguish percussion sounds from a recording of a musical performance [2]. Other researchers explored how to automatically organize music into genres using supervised learning methods [3]. Saunders utilized as features the average zero-crossing rate and the short time energy and applied tresholding method for distinguishing of speech and music from the radio broadcast [4]. Scheirer used different features in time, frequency and cepstrum domains and various classification methods to improve performance [5]. El-Maleh suggested a method for classification of audio signal into few categories [6]. Another paper describing content-based audio classification method based on neural network and genetic algorithm was published by Xi Shao et al. [7]. There were published a lot of classification approaches based on those algorithms which have several advantages with comparison to statistical classifiers. This sort of classifiers is strongly dependent on statistical data distribution while neural network classifiers are able to estimate nonlinear relationship between input and desired output data. Moreover, input data can be corrupted or incomplete; often the solution to a problem can be so complex that its description is not possible. In this case, using of artificial neural network is practically only one possibility. The most often used method of control classification by neural network is the multilayer Perceptron [8].

In this study, artificial neural network was applied in order to classify different audio classes especially from security and military areas, such as various shots and explosions.

2 Collecting data

At the beginning of this study, we had some audio data of real shots of various weapons but only few single shots of individual weapon. This number was not enough to train any effective classifier, so that we decided to create our own audio data.

Acquisition of real quality audio data for analysis consists in use of proper hardware configuration and mainly in abundant gunshots for individual weapons. While the former is easily performable, the latter represents more complex problem. It corresponds with acquisition of real shots in different surroundings and under various conditions (echo, reflections, humidity, and segmentation of terrain). We visited public shooting gallery where lot of gunshot series was taken. A few types of hand guns were tested (revolvers, colts, rifles), for example, Tokarev TT-33, STI-Spartan, CZ75SP-01, Winchester 1866, Arminius HW 38, complete list of tested guns is evident from Table 1 below.

Two microphones were used and located directly in shooting area in distance of 10 and 25 meters from firing position. Recorded gunshots were taken at sampling frequency 48 000 Hz and were free of noise but they contained echo. However, the most information which can be utilized in our analysis is presented in the first few milliseconds of the recorded sound.

3 Experimental

3.1. Sample processing

Different length of the raw sound was another aspect which we had to solve. For illustration, a raw waveform of a recorded audio signal, in this example, it is single gunshot from Arminius HW 38, is shown in Fig. 1. Samples were normalized in the way where amplitudes and frequencies of the waveform were preserved while their length became uniform. In addition, to keep compatibility with mentioned localization system, recorded sounds were resampled at 20 000 Hz. This process can be divided into these steps:

- Separation of individual gunshots
- Resampling at specific frequency
- Direct component removal
- Recalculation samples to acoustic pressure

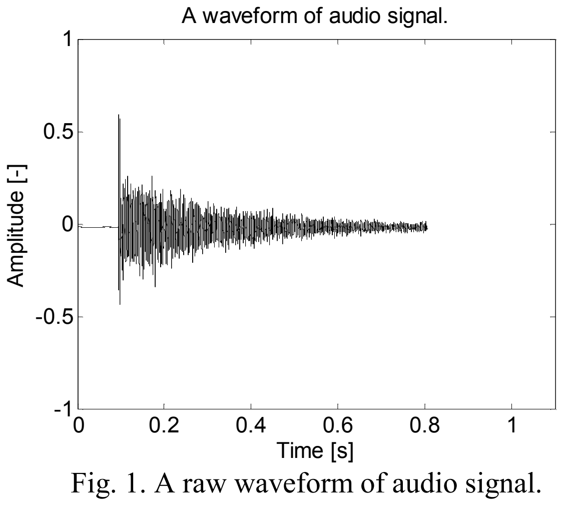

Fig. 1. A raw waveform of audio signal.

As a result of separation of individual gunshots, 100 milliseconds of rest audio segment are kept before each shot due to not cutting of the shot beginning. Individual samples were recalculated to acoustic pressure using

$$y_{pa}(k) = y(k) \cdot \frac{U_r}{C_{mic}} \tag{1}$$

where $y(k)$ is recorded sound signal, its values are from interval <-1; 1>, U_r is voltage range of the sampling unit, C_{mic} is sensitivity of used microphone (depends on manufacturer, in our case C_{mic}=31,6.10^{-3} V.Pa^{-1}).

All samples were normalized to reference distance 10 m due to their better mutual comparison according to:

$$y_{pa10m}(k) = y_{pa}(k) \cdot \frac{d_{mic}}{d_{ref}} \tag{2}$$

where d_{mic} is distance of the microphone form audio source and d_{ref} is reference distance. Recalculated audio signal is depicted in Fig. 2.

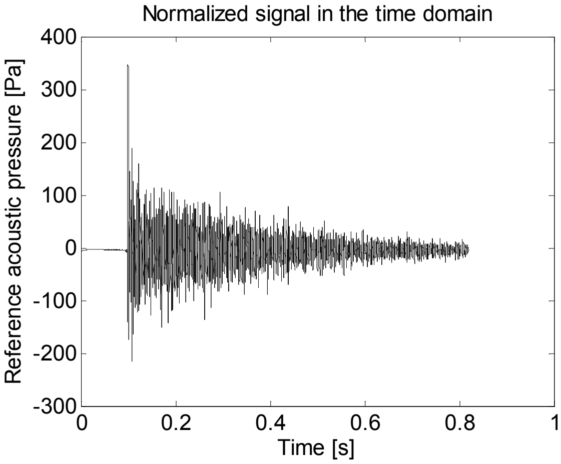

Fig. 2. Normalized audio signal in time domain.

Acoustic pressure level can be calculated using

$$L_P = 20 \cdot \log_{10}\left(\frac{p_a}{p_0}\right) \tag{3}$$

where p_a is measured value of acoustic pressure, p_0 is limit the audibility 2.10^{-5} Pa (corresponds to 0 dB).

3.2. Frequency analysis

Fourier analysis is a set of mathematical techniques, all based on decomposing signals into sinusoids.

Signals are very often converted from time domain to the frequency domain through the Fourier transform. It converts the information about signal to a magnitude and phase component for each frequency [9]. Another quantity describing the signal in frequency domain is power spectrum density (PSD). It describes how the energy of a signal is distributed along frequency axis. Due to spectrum study it can be determined which frequencies are contained in the signal and which are not [10]. The Fourier transform $X(\omega)$ of the input time-continuous signal $x(t)$ is defined as follows:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \qquad (4)$$

where ω is angular frequency, j is complex unit, t is time.

As we process audio signals which are digitized and are not infinite, it is necessary to compute the discrete Fourier transform (DFT) [11]. The DFT replaces the infinite integral with a finite sum

$$X(\omega_k) = \sum_{n=0}^{N-1} x(t_n) e^{-j\omega_k t_n} \qquad (5)$$

where $k = 1, 2 \ldots N-1$, t_n is nth sampling instant, N is number of samples, ω_k is kth frequency sample. Calculated one-sided power spectral density of mentioned single gunshot from Arminius HW 38 is shown in Fig. 3.

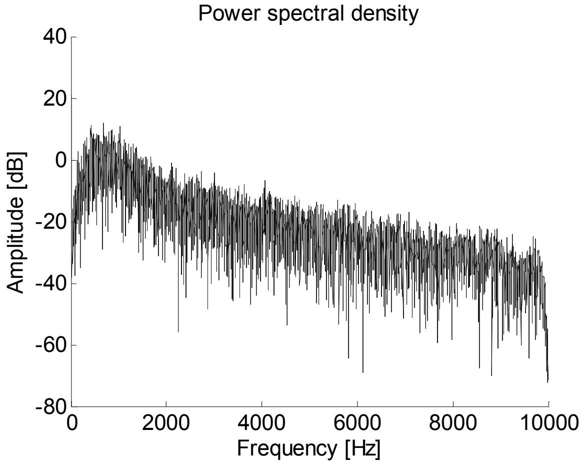

Fig. 3. Power spectral density of the signal.

For frequency analysis the first 0.1 s was omitted and the signal was taken of length 200 milliseconds.

3.3. Feature vector selection

Studying frequency analysis of recorded audio signals taken in real environment, four types of feature vectors were selected as possible candidates. All of them were derived from power spectral density. A one-sided PSD contains the total power of the signal in the frequency interval from DC component to half of the Nyquist frequency, which was in our case 10 kHz. However, for analysis, frequency interval was chosen up to 5 kHz due to microphone frequency characteristic and elimination of overtones, see Fig. 4.

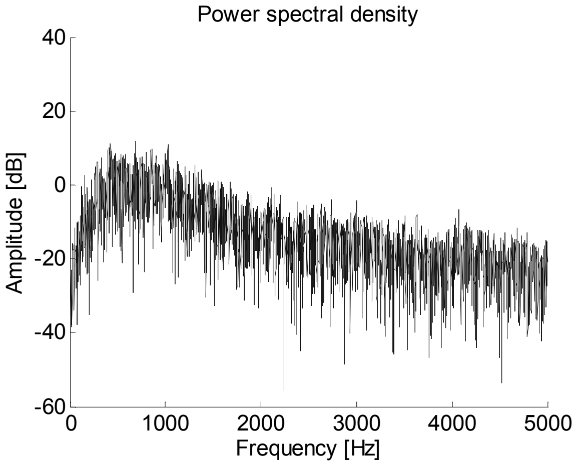

Fig. 4. Cut power spectral density (5 kHz)

3.3.1 Reduced PSD

Full PSD contains ten thousand samples, after elimination we get half but it is still big number. For that reason, we define reduction factor R which reduces number of samples by taking only (kR)th samples where $k=1\ldots N/R-1$. As a result of this reduction, R times less number of samples (frequency scale is maintained) is taken as feature vector, see Fig. 5.

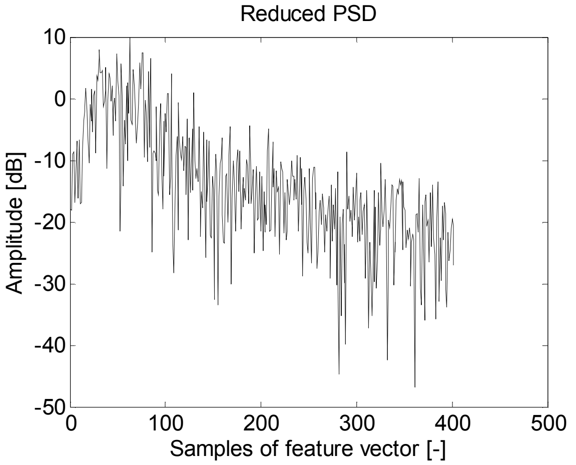

Fig. 5. Reduced power spectral density

3.3.2 Significant points of moving averaged PSD

Only a few significant points (the most powerful components including frequency information) of moving averaged PSD were taken in order to make feature vector as small as possible and simultaneously to keep information value, these points are depicted by circles in Fig. 6.

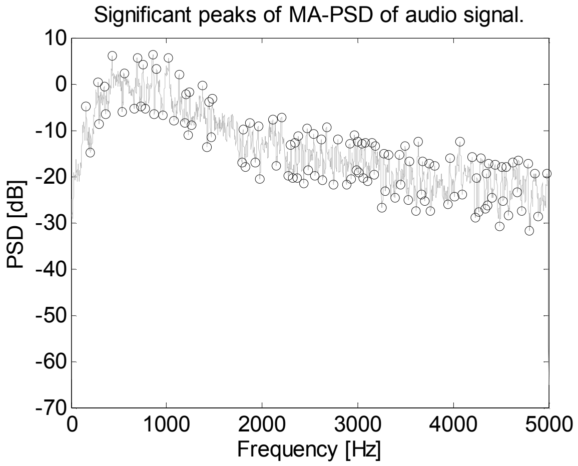

Fig. 6. Significant peaks of moving averaged PSD

3.3.3 Autocorrelation function of PSD

This characteristics represents normed autocorrelation function of power spectral density which is smoothed with moving average and then the number of points is reduced by a factor, see Fig. 7.

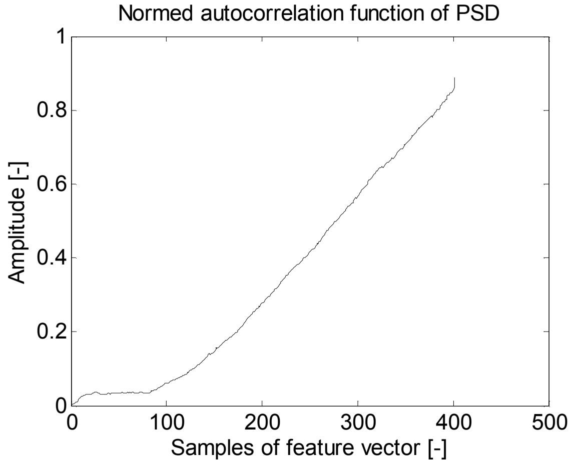

Fig. 7. Normed autocorrelation function of power spectral density.

3.3.4 Spectrogram

The spectrogram is a specific representation of spectrum (frequencies and amplitudes) as a function of time. It is essential tool of frequency analysis applicable in many fields, especially in audio signal processing. The spectrogram can be defined as an intensity plot of the Short-Time Fourier Transform (STFT) magnitude. The STFT is a sequence of FFTs of windowed data blocks, where the windows are overlap in time [12], [13]. Typical values are around 25-50 %, in our case, due to relative short duration of analysed sounds, we used overlapping up to 85 %. These higher values results in better resolution of the spectrogram, on the other hand, it is more time-demanding so that there is a need to find a compromise.

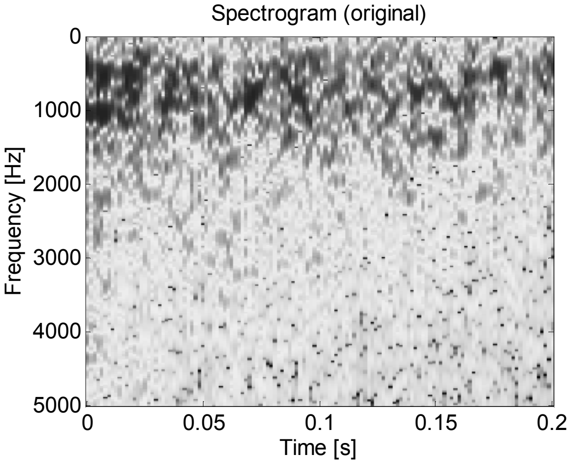

Fig. 8. Original spectrogram

For feature vector constitution there was a need to decrease resolution (due to length of the vector). Original spectrogram was modified with down-sampling, see Fig. 9 and then feature vector was arranged, as can be shown in Fig. 10.

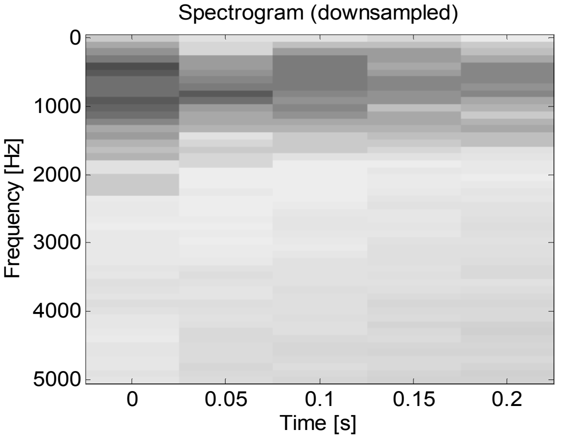

Fig. 9. Down-sampled spectrogram

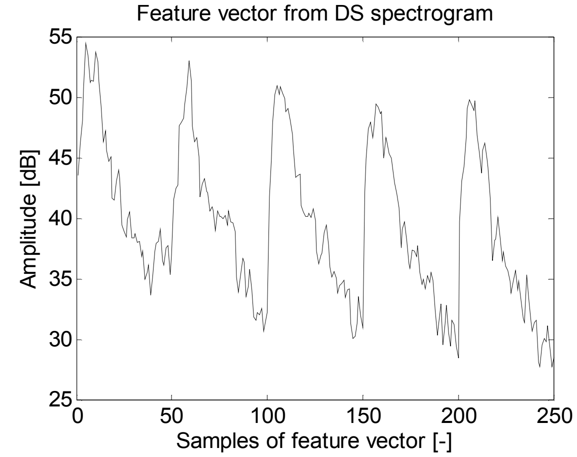

Fig. 10. Feature vector taken from down-sampled spectrogram.

3.4. Software implementation

Software application for audio signal classification was created in Matlab software with utilization of graphical user interface (GUI). ANN was implemented using neural network toolbox; Levenberg-Marquardt back-propagation algorithm and gradient descent with momentum weight/bias learning function were used. Mean squared error (MSE) was chosen as performance function. The application consists of two main parts; the first is designed for training of artificial neural network while the second is focused on classification of audio samples.

3.5. Patterns preparation

Available recorded samples were processed, our first approach meant dividing the samples into ten groups according to handgun type; their complete list is mentioned in Table 1. Moreover, in the same way, some of available recorded samples were divided into three classification groups according to gun category (Table 2). In both cases, samples were further split into learning and testing patterns.

Table 1. Various handguns, 10 classification groups

	Gun type	*Number of patterns for*	
		learning	*testing*
1.	Norinco W97 12	10	20
2.	Colt 1873	10	20
3.	Saiga 308	10	20
4.	Tokarev TT-33	10	20
5.	STI-Spartan	10	20
6.	Arminius HW 38	10	20
7.	CZ75SP-01	10	20
8.	S&W 629	10	20
9.	S&W 27	10	20
10.	Winchester 1866	10	20
	Total:	***100***	***200***

Table 2. Gun categories, 3 classification groups

	Gun category	*Number of patterns for*	
		learning	*testing*
1.	Handguns	50	100
2.	Big bore guns	10	20
3.	Explosions	5	10
	Total:	***65***	***130***

4 Results and discussion

Artificial neural network was repeatedly learned under various conditions, especially structure of network and its parameters, feature vector settings and other. The best achieved results are summarized in Table 3, where there are five columns. *Feature vector* represents input into the network (FV1, FV2, FV3 and FV4). *HN* means number of neurons in hidden layer. *FV length* indicates number of neurons in input layer and depends on feature vector settings. *FV time* is duration of feature vector computation of one sample. *MSE* is performance function computed after 100 iterations.

Table 3. Parameters of neural network and achieved test results.

Feature vector		*HN [-]*	*FV length [-]*	*FV time [s]*	*MSE [-]*
FV1	Reduced PSD (RF = 10)	4	402	0.02	10^{-5}
FV2	Significant points of MA-PSD (11 points avg., taken 30 sign. points)	4	60	0.03	0.22
FV3	Autocorrelation function of PSD	4	402	0.05	0.082
FV4	Spectrogram (0.2 s)	6	250	4.5	0.15

The most important parameter is success classification rate (SCR) which is ANN successfulness in correct classification, see Table 4.

Table 4. The best obtained results of the classification process.

	SCR			
	FV1	*FV2*	*FV3*	*FV4*
10 groups	42 %	38 %	59 %	46 %
3 groups	95 %	85 %	96 %	94 %

According to reduction degree of information from frequency analysis we obtained results which correspond to that consideration. The best performance had feature vector containing the most number of information. The more reduction is used the less ability of the neural network to learn and correctly classify.

5 Conclusion

In this paper, classification of audio sources using artificial neural network was presented. Success classification rate lies in the convenient discrimination of the feature vector in the feature vector space. Four types of characteristics based on frequency analysis were chosen and gradually used as inputs for neural network. Using of autocorrelation function of power spectral density as a feature vector proved the best performance for the classification. During classification into three groups we achieved the best success classification rate 96 %; into ten groups we achieved the best success classification rate 59 %.

The frequency features utilized in this work could be improved if they give preference to information near the beginning of the wave instead of regarding all time windows with uniform importance. Problem of the classification of audio sources which is supposed to be similar to each other is deepened especially in various surrounding. Audio signal propagation is different in open space, inside built-up areas, hilly terrain or in the forest. Another difficulty consists in echo and consequent explosions.

SCR could be increased via greater number of training patterns but this corresponds to problematic acquisition of real shots and explosions in different surroundings and conditions.

All of these mentioned aspects are currently solved and further study is prepared.

6 Acknowledgement

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic under the Research Plan No. MSM 7088352102 and by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

References:

[1] P. Dostálek, V. Vašek, V. Křesálek, M. Navrátil, *Utilization of Audio Source Localization in Security Systems*, In Proc. ICCST, 2009, pp. 305-311.

[2] V. Sandvold, F. Gouyon, and P. Herrera, *Drum sound classification in polyphonic audio recordings using localized sound models*, In Proc. ISMIR, 2004.

[3] K. West and P. Lamere, *A model-based approach to constructing music similarity functions*, EURASIP J. Appl. Signal Process., 2007, pp.149-149

[4] J. Saunders, *Real-time Discrimination of Broadcast Speech/Music*, In Proc. ICASSP-96, 1996, pp.993-996.

[5] E. Scheirer and M. Slaney, *Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator*, In Proc. ICASSP97, Vol.2, 1997, pp.1331-1334.

[6] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal, *Speech/Music Discrimination for Multimedia Application*, In Proc. ICASSP00, 2000.

[7] X. Shao, X. Changsheng, X. and M. S. Kankanhalli, *Applying Neural Network on the Content-Based Audio Classification*, In Proc. ICICS-PCM, 2003

[8] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1998, Prentice Hall.

[9] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, second edition. California Technical Publishing, 1999.

[10] M. J. Lighthill, *Introduction to Fourier Analysis*, Cambridge University Press, 1958.

[11] J. O. Smith, *Mathematics of the Discrete Fourier Transform (DFT), with Audio Applications*, Second Edition, http://ccrma.stanford.edu/~jos/mdft/, Apr. 2007, online book

[12] A.V. Oppenheim, and R.W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989, pp.713-718.

[13] P. Duhamel, and M. Vetterli, *Fast Fourier Transforms: A Tutorial Review and a State of the Art*, Signal Processing, Vol. 19, April 1990, pp. 259-299.