International Journal of Imaging and Robotics,
[Formerly known as the "International Journal of Imaging" (ISSN 0974-0627)]
Volume 16; Issue No. 3; Year 2016; Int. J. Imag. Robot.
ISSN 2231–525X; Copyright © 2016 by IJIR (CESER PUBLICATIONS)

Sound Identification Model in Terms of the Active Perception Theory

Vasiliy E. Gai and Igor V. Polyakov

Institute of Radioelectronics and Information Technologies
Nizhny Novgorod State Technical Universtiy n.a. R.E. Alekseev
K. Minina, 24, Nizhny Novgorod, Russian Federation, 603950
vasiliy.gai@gmail.com, polyakovigor92@gmail.com

ABSTRACT

The paper describes the solution to the problem of sound description choice for its compact presentation (production of a signal digital fingerprint), as well as the algorithm of information retrieval in a digital fingerprint database. The obtained presentation and retrieval algorithm can be used in development of music identification systems, and radio and TV broadcast monitoring (for example, in order to monitor commercials broadcasting).

Keywords: sound identification system, active perception theory.

2010 Mathematics Subject Classification: 68T10, 93E12, 94A13.
1998 Computing Classification System: C.3, H.5.5, I.5.4.

1 Overview of Current Sound Identification Systems and Algorithms

1.1 Identification Purpose Description

The amount of records stored currently in the Internet is huge. For example, the Yandex.Music service stores about five million records, the Shazam service – five billions. Its evident that such a situation requires the option of quick and exact search among available musical records. Moreover, with the increase of the record number, the importance of this option will rise.
The current sound retrieval algorithms can be divided into two trends:

1. Content-Based Sound Retrieval, CBSR;
2. Description-Based Sound Retrieval, DBSR.

Description-based sound retrieval has a number of drawbacks:

1. textual description of each sound shall be created for the search, which is troublesome;
2. an ambiguous correspondence between the content and textual description of the sound file lowers precision and completeness of retrieval.

As opposed to the CBSR systems, the content-based sound retrieval systems do not require any additional file information. The search is performed on the basis of analysis and comparison of sound file specifications.
A sound identification system should quickly and correctly find a sound in a data base using its fragment. A signal fragment may be corrupted, and its duration can be just a few seconds. Ideally, the algorithm operation time should not depend on the number of records in the database (Kamaladhas and Dialin, 2014).
This paper is focused on the application of the active perception theory (APT) and sound identification problem solution (Gai, 2015).

1.2 Identification Algorithm Overview

The simplest solution to the problem of sound identification by its fragment consists in direct comparison of the target and initial signal magnitude in a definite moment of time. The drawbacks of this algorithm are low resistance to target signal distortion, as well as long algorithm operation time.
It has been established that to build up a jamproof sound identification system, a signal description shall be created (signal digital fingerprint). Ideally, the digital fingerprint shall be resistant to various signal distortions.
In this case, fragment-based sound identification consists in the following:

1. a fingerprint is extracted from the reference signal and stored in a data repository;
2. the unknown signal fingerprint is compared to a great amount of database fingerprints;
3. potentially compliant signals are checked for perfect match.

Lets consider some well-known digital fingerprint generation algorithms. The digital fingerprint generation algorithm described in (Haitsma and Kalker, 2002) is as follows:

1. a 3-second sound is divided into segments overlapping in time;
2. a 32-bit digital fingerprint is generated for each sound fragment in the form of the Fourier transform.

Within this algorithm two three-second sounds are deemed coinciding, if the Hamming distance between them is less than a certain threshold value. Algorithm advantages: high resistance to interference and operation speed, drawbacks: long fingerprint generation time due to segment overlapping, lack of signal compression / extension invariance.
The algorithm described in (Wang et al., 2003) is as follows. A signal spectrogram is generated basing on the Fourier transform. A signal fingerprint is calculated basing on the points corresponding to the local spectrogram maxima. Then, to improve description resistance to interference, links are created between the local maxima pairs. The obtained link information is used for generation of a hash code packed into a 32-bit value. As a result, a digital fingerprint is made basing on the multitude of such hash codes and used for sound identification.
In papers (Kurth, 2002; Poulos, Deliyannis and Floros, 2012) the windowed Fourier transform with various window overlapping value and degree is used for obtaining a signal spectrogram.

Basing on the obtained spectrogram stable signal features are calculated and its digital fingerprint is generated.
There are algorithms using wavelet transforms for digital fingerprint generation (Baluja and Covell, 2007).
Sound identification approach using Guassian mixture models, as well as based on other features, such as the Shannon entropy, the Renyi entropy, and mel-cepstral coefficients is described in (Ramalingam and Krishnan, 2006).
Paper (Mitrović, Zeppelzauer and Breiteneder, 2010) describes most well-known features calculated basing on a sound. Content-based sound retrieval systems architecture is described, as well as the problems connected with designing such systems and solution methods.
The developers of some sound identification systems often use algorithms and methods from machine learning domain for validation of various sound feature sets and automatic selection of the most efficient ones.

2 Sound Identification in Terms of the Active Perception Theory

2.1 Information Model of Sound Identification System

Information transformations of a signal for sound identification purposes, in terms of the active perception theory, are shown in fig. 1.

2.2 Sound Description Generation Algorithm

Digital fingerprint generation algorithm comprises the following stages:

1. division of sound S into equal segments with overlapping O: $s = \{s_m\}$, $m = \overline{1, M}$, M is the number of sound segments S (O is segment overlapping coefficient);

2. obtaining sound spectrum using the U-transform (Utrobin, 2004):

$$d = U[s_m], \tag{2.1}$$

   where U is U-transform calculation statement, $d = \{d_m\}$, $m = \overline{1, M}$;

3. screening out segments, which are not resistant to signal distortion:

$$d' = Q[d], \tag{2.2}$$

   where Q is distortion-resistant signal segment select statement, $d' = \{d'_n\}$, $n = \overline{1, N}$, N is the number of stable segments, $N \leq M$;

4. closed group generation on the basis of stable segment spectra:

$$p = P[d'], p = \{p_n\}, n = \overline{1, N}, \tag{2.3}$$

   where P is closed group calculation statement.

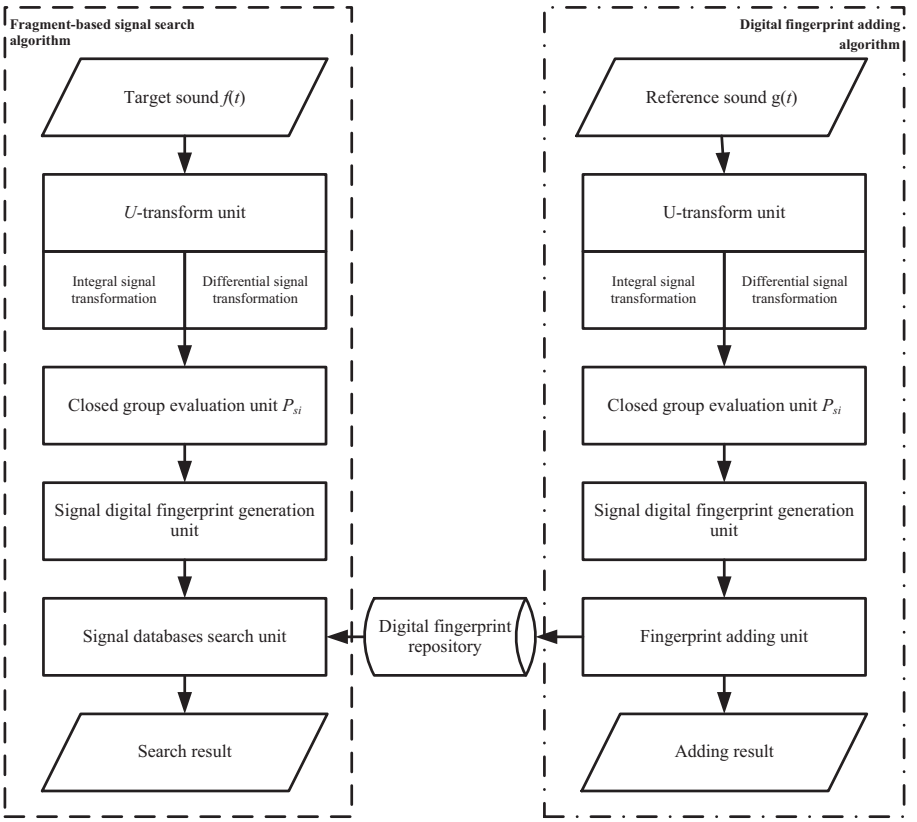

Figure 1: Information model of sound identification system

5. digital fingerprint generation for stable sound segments:

$$q = Q[p, K], q = q_n, n = \overline{1, N}, \tag{2.4}$$

where Q is fingerprint calculation statement, K is fingerprint length (in characters).

Let's take a detailed look at the fingerprint generation process. In the APT statements, complete and closed groups can be used for creating description. It has been established empirically that closed groups are most resistant to interference (Gai, 2014). If the sufficient number of closed group images is not generated for any input signal segment, such segment is considered nonsignificant, and no digital fingerprint is generated for it. Closed group mass is calculated for each image. Closed group mass is connected with group resistance to signal distortion: the greater the mass of the group calculated basing on a certain signal, the greater is the probability that such group will be included in the signal description after its distortion. Therefore, the first K-s of closed groups selected on the basis of their mass are used for digital fingerprint generation.

Closed group image consists of 16 logical values. Such image can be presented as one Unicode character. As a result of digital fingerprint generation a line consisting of the K characters is generated. Any decrease in the K value will result in the increase in the number of matching characters.

The obtained digital fingerprint together with other segment data (identifier, name, track author, segment identifier, and digital fingerprint identifier) is stored in the database. This data is used for subsequent fragment-based sound search.

2.3 Sound Database Search Algorithm

Sound search algorithm compares input signal fingerprints to the fingerprints of signals stored in the database. This algorithm, as it pertains to digital fingerprint generation, corresponds to the one described in paragraph 2.2.

After J fingerprints are generated, they are directed to the database for the search of identical fingerprints using the SQL request. The next set of digital fingerprints is generated before database response is received.

If the target fingerprint is not found, it is added to the temporary table together with the segment identifier, signal identifier and next two segments and their digital fingerprints. Fingerprint search in a database is performed using an index created in accordance with the fingerprint line field. After searching through all input signal fingerprints the obtained temporary table is returned from the repository.

The generated target signal fingerprints are compared with those from the temporary table following the matching ones. Three successful comparisons of digital fingerprints are enough to determine unambiguously that the input signal matches a signal in the database. If they match, it means that the signal is found, otherwise the search continues. If no matches are found, it is considered that the signal is not present in the database. Fig. 2 shows the sound search flow diagram.

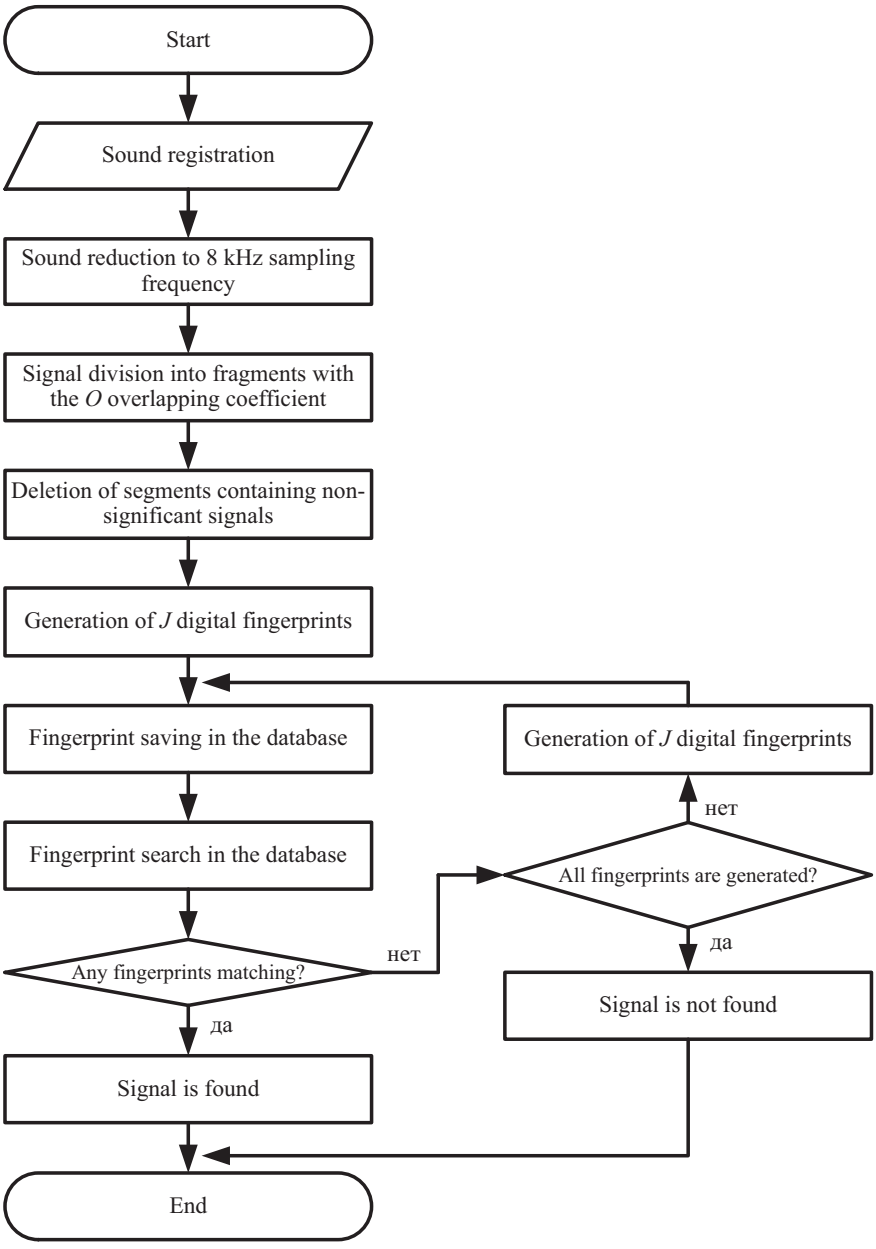

Figure 2: Sound search flow diagram

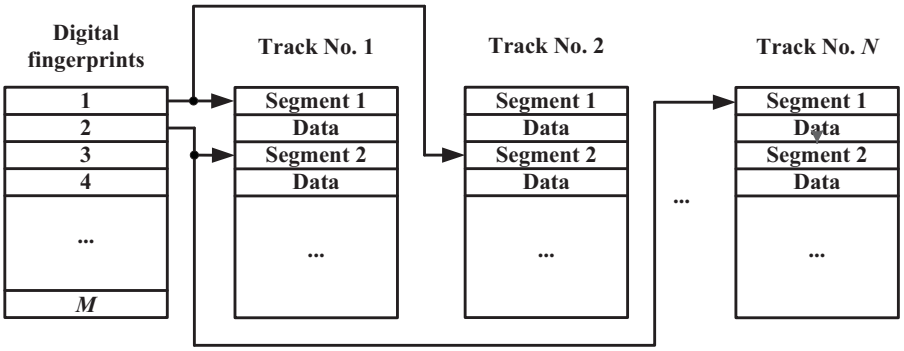

Figure 3: Connection between digital fingerprints and signal segments

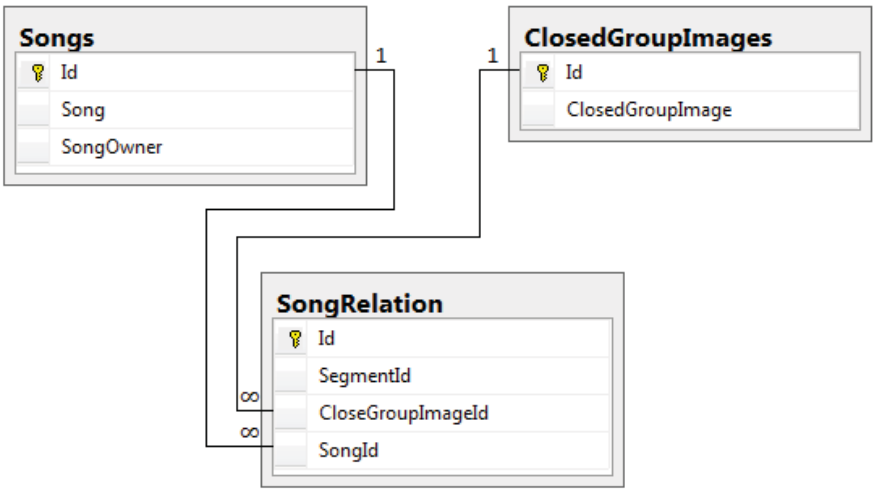

Figure 4: Database structure

2.4 Database Structure

Digital fingerprints and music record data are stored in a relational database. A digital fingerprint of a signal segment is not segment's unique identifier, therefore one fingerprint can correspond to several segments of one or several sounds (see fig. 3). This digital fingerprint property is reflected in the database structure.

Fig. 4 shows database structure that meets the above requirements.

Let's consider the tables included into database and their fields. The Songs table contains sound description:

1. Id (integer) is signal identifier (table primary key);
2. Song (string) is signal name;
3. SongOwner (string) is performer.

The ClosedGroupImages table contains the images (digital fingerprints) of signal segments:

1. Id (integer) is image identifier (table primary key);
2. ClosedGroupImage (string) is sound segment image.

SongRelation is an auxiliary table designed for connection of the signal segment image table (ClosedGroupImages) with the one containing signal information (Songs):

1. Id (integer) is record identifier (table primary key);
2. SegmentId (integer) is the number of the segment in a signal;
3. CloseGroupImageId (integer) is identifier of an image from the ClosedGroupImages table;
4. SongId is identifier of a signal from the Songs table.

Apart from the tables, the database contains indices for B-tree search facilitation in the following fields:

1. ClosedGroupImages.ClosedGroupImage – unclustered;
2. ClosedGroupImages.Id – clustered, primary key-based;
3. SongRelation.SongId – unclustered;
4. SongRelation.CloseGroupImageId – unclustered;
5. SongRelation.Id – clustered, primary key-based.

Taking into account that the indices are based on a B-tree, the average algorithmic complexity of searching an element in a database is equal to the medium complexity of element search in the B-tree $O(log(N))$, where N is the number of records in the table. Database interfacing is performed by sending SQL-requests and selection of stored procedures.

3 Computational Experiment

Let's test the developed sound identification system. Table 1 contains the data on musical records used in the computational experiment. The proposed system and the known will be compared by a number a criteria:

1. digital fingerprint size;
2. record identification time;

Table 1: Records used in the experiment

No.	Name	Author	Genre	Duration (in secs.)
1	Coco Jambo	Mr. President	Pop	219
2	Ohne dich	Rammstein	Industrial metal	270
3	The winter	Antonio Vivaldi	Classic	213
4	You don't know	50 Cent, Eminem	Rap	258
5	I was made for loving you	Scooter	Hard rock	212

3. signal distortion resistance;

4. classification precision.

In the computation experiment the signal segment length at digital fingerprint generation was taken equal to 64 counts. The length of signal segment description was taken equal to 7 characters ($K = 7$). For generation of a digital fingerprint of the signal identified the $O = 63/64$ overlapping coefficient was applied.
As of the date of the research, the database contained information on 1000 music tracks. In the modern systems the number of records is by several orders more (millions of records). It should be taken into account when analyzing the research results.
System testing was performed on an N76M Asus laptop, with Intel Core i7 processor, 2 cores of 2.40 GHz frequency, 8 GB random-access memory, Windows 8.1 operating system, Microsoft SQL Server 2012 Express Edition DBMS. Mobile client-side programs were run on Samsung Galaxy S3 smartphone, with Samsung Exynos 4412 processor, 1400 MHz, 8 GB random-access memory, Android 4.4.2 Kitkat operating system. No data on other server-side identification system computational resources has been found.

3.1 Digital Fingerprint Size

Let's determine the size of a fingerprint generated per one second of a sound. For the system under consideration, the one-second fingerprint size may vary, but the average values for different records are similar. At the average, 16.2 significant segments fall on one signal second, and a 14-byte digital fingerprint corresponds to each of them. Thus, in the system developed a 226.8-byte fingerprint is generated per one signal second at the 8 kHz sampling frequency.
It is of importance that the number of segments cast out at the stage of digital fingerprint generation is rather high. In practice it amounts to 80-85% of all sound segments. Nevertheless, it does not impede the further search, since the rest of the segments is distributed equally throughout the signal and is most resistant to noise.
In article (Haitsma and Kalker, 2002) it is shown that the used algorithm of digital fingerprint generation provides for obtaining one sub-fingerprint for each 11.6 (ms) of a sound, sub-fingerprint size being equal to 32 bits, which constitutes 344.8 bytes per second.
In algorithm description in (Wang et al., 2003) it is stated that the average number of sub-fingerprints per sound second is equal to 10, sub-fingerprint size being equal to 64 bits, i.e. a

Table 2: Record identification duration (in seconds)

No.	Developed system	Tunatic	Echoprint	Shazam	Soundhound
1	3.2	15.3	7.3	5.1	8.5
2	4.4	9.4	12.5	6.2	12
3	4.7	13.9	8.9	6.3	6.8
4	2.2	10.2	10.5	3.2	6.5
5	3.6	12.7	10.7	9.3	5.9
Average	4.0	12.3	9.9	6.0	7.9

80-byte fingerprint is generated for one signal second. See the obtained values in table 4.

3.2 Record Identification Time

Record identification time depends on the noise level and the popularity of the target record, since the data on the popular record may be stored in the frame memory, increasing the speed of retrieval. The research results are given in table 2. During the research the Internet reception speed was 10.15 Mbps, and transmission speed – 15.27 Mbps.
The developed system showed the best results by the ”identification time” criterion. This result was considerably influenced by a small database volume, and the fact that, as opposed to other systems, the calculations were carried out on one computer (without Internet connection). Shazam showed the best results among the other systems.

3.3 Signal Distortion Resistance

Let’s consider identification system resistance to signal distortion caused by noise. In Table 3 for each system the minimum signal/noise ratio for correct signal identification is specified. The experiment was carried out for white, pink and red noise.
It is apparent from Table 3 that the developed system is inferior to the modern identification systems in terms of resistance to signal distortion by noise. In general, the identification systems better cope with red and pink noise, than white noise.

3.4 Classification Precision

As a result of the developed system study it has been established that type I error is equal to 0.002, and type II error is equal to 0.005 (for the database composed of 1000 records).

3.5 Research Result Summary

The research results are given in Table 4. A dash in the table means that the parameter value has not been determined.

Table 3: Noise masking experiment results

Noise type	No.	Developed system	Tunatic	Echoprint	Shazam	Soundhound
White	1	11.4	9.9	10.2	4.7	5.1
White	2	11.2	10.2	10.2	4.7	5.0
White	3	11.5	10.8	10.2	4.7	5.1
White	4	11.5	9.7	10.2	4.7	5.2
White	5	11.2	9.9	10.2	4.7	5.1
Pink	1	10.9	4.2	4.6	4.2	2.2
Pink	2	10.6	4.2	5.5	4.2	2.2
Pink	3	11.7	6.7	5.5	4.2	2.2
Pink	4	11.4	4.2	5.5	4.2	2.2
Pink	5	11.1	4.2	5.5	4.2	2.2
Red	1	11.4	1.8	2.6	5.4	5.7
Red	2	11.4	1.8	2.2	5.4	5.7
Red	3	11.4	2.0	2.4	5.4	5.7
Red	4	11.2	2.2	2.5	5.4	5.7
Red	5	11.4	1.5	2.6	5.4	5.7

Table 4: Summary table of comparison results

Comparison criterion	Developed system	Tunatic	Echoprint	Shazam	Sounhound
Size of digital fingerprints per signal second (in bytes)	226.8	-	344.8	80	-
Average record identification time (in seconds)	4.022	12.3	9.98	6.02	7.94
Signal-to-noise ratio for white noise	11.2	9.7	10.2	0.97	5.0
Signal-to-noise ratio for pink noise	10.6	4.2	4.6	4.7	2.2
Signal-to-noise ratio for red noise	11.4	1.5	2.2	4.2	5.7
Minimum signal fragment duration (in seconds)	0.4	6	4.5	5.4	3
Time of record addition to the database (in seconds)	1.95	-	-	-	-

Conclusion

A sound identification system information model is proposed in the article. The created system, as opposed to the previous ones, provides for identifying a sound with less than one second duration.

The results of comparison of requirements to the minimum signal fragment duration suggest that the developed system can successfully solve the problem of identifying a signal fragment with less than one second duration (at a low noise level). A short time of adding a one-minute record to the database enables to quickly increase the number of records in the database, and a small size of the digital print provides for the storage of most popular record fingerprints in the frame memory (randomizing option), facilitating their retrieval. The data on about 79 thousand 60-second sounds can be stored in 4 GB frame memory.

Thus, the following results have been obtained:

1. an algorithm for compact digital signal presentation has been developed;
2. an algorithm for data retrieval in a digital fingerprint database has been developed.
3. the possibility of applying the active perception theory to solution of the digital track generation problem has been proved.

The practical relevance consists in the development of a music identification system based on the proposed model.

Basing on the proposed model, sound identification software has been developed. Pilot study of the developed sound identification system has been carried out, as well as its comparison with the modern identification systems. The developed system can compete favorably with the current systems in terms of digital fingerprint size per signal second, signal identification time, and minimum recognizable signal fragment duration.

References

Baluja, S. and Covell, M. 2007. Audio fingerprinting: Combining computer vision & data stream processing, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 2, IEEE, pp. II–213.

Gai, V. E. 2014. A study of the stability of sound signal description, *Pattern Recognition and Image Analysis* **24**(4): 463–466.

Gai, V. E. 2015. The method and algorithmic support of the determining the presence of information message in a priori an uncertain sound signal vasiliy e. gai, *International Journal of Tomography & Simulation* **28**(2): 1–14.

Haitsma, J. and Kalker, T. 2002. A highly robust audio fingerprinting system., *ISMIR*, Vol. 2002, pp. 107–115.

Kamaladhas, M. D. and Dialin, M. M. 2014. Robust fingerprint extraction algorithm for audio identification, *International Journal of Tomography & Simulation* **26**(2): 55–62.

Kurth, F. 2002. A ranking technique for fast audio identification, *Multimedia Signal Processing, 2002 IEEE Workshop on*, IEEE, pp. 186–189.

Mitrović, D., Zeppelzauer, M. and Breiteneder, C. 2010. Features for content-based audio retrieval, *Advances in computers* **78**: 71–150.

Poulos, M., Deliyannis, I. and Floros, A. 2012. Audio fingerprint extraction using an adapted computational geometry algorithm, *Computer and Information Science* **5**(6): p88.

Ramalingam, A. and Krishnan, S. 2006. Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting, *Information Forensics and Security, IEEE Transactions on* **1**(4): 457–463.

Utrobin, V. A. 2004. Physical interpretation of the elements of image algebra, *Physics-Uspekhi* **47**(10): 1017–1032.

Wang, A. et al. 2003. An industrial strength audio search algorithm., *ISMIR*, pp. 7–13.