Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Zvezek A

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Volume A

Uredili / Edited by

Marko Bohanec, Matjaž Gams, Vladislav Rajkovič, Tanja Urbančič, Mojca Bernik,
Dunja Mladenić, Marko Grobelnik, Marjan Heričko, Urban Kordeš, Olga Markič,
Jadran Lenarčič, Leon Žlajpah, Andrej Gams, Andrej Brodnik

11.–15. oktober 2010 / October 11th–15th, 2010
Ljubljana, Slovenia

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Zvezek A

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Volume A

Inteligentni sistemi
Vzgoja in izobraževanje v informacijski družbi
Izkopavanje znanja in podatkovna skladišča (SiKDD 2010)
Sodelovanje, programi in storitve v informacijski družbi
Kognitivne znanosti
Robotika
MATCOS 2010 (Mini-konferenca v uporabnem teoretičnem računalništvu)

Intelligent Systems
Education in Information Society
Data Mining and Data Warehouses (SiKDD 2010)
Collaboration, Software and Services in Information Society
Cognitive Sciences
Robotics
MATCOS 2010 (Mini-conference on Applied Theoretical Computer Science)

Uredili / Edited by

Marko Bohanec, Matjaž Gams, Vladislav Rajkovič, Tanja Urbančič, Mojca Bernik, Dunja Mladenić, Marko Grobelnik, Marjan Heričko, Urban Kordeš, Olga Markič, Jadran Lenarčič, Leon Žlajpah, Andrej Gams, Andrej Brodnik

http://is.ijs.si

11.–15. oktober 2010 / October 11th–15th, 2010
Ljubljana, Slovenia

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Izrael
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Karl Pribram, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Finland
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Miklós Krész, Hungary
József Békési, Hungary

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek, co-chair
Lana Jelenkovič
Jana Krivec
Mitja Lasič

Programme Committee

Franc Solina, chair
Viljan Mahnič, co-chair
Cene Bavec, co-chair
Tomaž Kalin, co-chair
Jozsef Györkös, co-chair
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Marjan Pivka
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Tomaž Šef
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
David B. Vodušek
Baldomir Zajc
Blaž Zupan
Boris Žemva
Janez Žibert
Leon Žlajpah

RECOGNITION OF INDIVIDUAL ANIMALS BY THEIR VOCALIZATION

Ivan Kraljevski[1], Solza Grceva[1], Igor Stojanovic[2], Zoran Gacovski[1], Biljana Spireva[1]
[1]Faculty for ICT, FON University, bul. Vojvodina bb, Skopje, Macedonia
[2]Faculty of Computer Science, University Goce Delcev, Toso Arsov 14, 2000 Stip, Macedonia
E-mail: {ivan.kraljevski, solza.grceva, zoran.gacovski, biljana.spireva}@fon.edu.mk,
igor.stojanovik@ugd.edu.mk

ABSTRACT

The recent advance in development of digital signal processing (DSP) and analysis techniques in human speech recognition systems has very significant influence on the field of bioacoustical research on animals [1]. In this paper, we present our approach in recognition of animal vocalizations with techniques for voice recognition. Vocalizations induced by physiological characteristics of the vocal tract of individual animals were encoded by Time Encoded Signals (TES). Further on, Artificial Neural Networks (ANN) was used for voice samples classification. The main benefit of the proposed technique is the application of such systems on low complexity DSP hardware due to low processing requirements [2]. In this paper, a system for individual "speaker" recognition of dogs barking samples was created and evaluated.

1 INTRODUCTION

During evolution many animal species created some complex forms of communications by vocalization. By that, the animals usually express their internal physiological state (hunger, thirst, matting, anxiety, etc.) as well as particular organism's condition influenced by various environmental factors (temperature, atmospheric pressure). By appropriate analysis of animal's vocalization, particular anatomical characteristics could be estimated. For example, the vocal tract length (VTL) is specific for different species and breeds, even for individual animals. The standard speaker and speech recognition techniques implemented in voice driven applications for individual animal recognition by their vocalization, is based on the assumption that the most mammal vocalizations (but not all) match frequency ranges of human vocalization [1].

This paper outlines a system for automatic individual "speaker" recognition of dogs by their barking samples, based on anatomical characteristics related with particular vocalization. The system uses Time Encoded Signals (TES) and Artificial Neural Networks (ANN) for classification of individual animals.

2 METHODS

There are many acoustical and anatomical similarities in the process of vocal sounds production - vocalization in mammals. The primary acoustical signal is created in the vocal folds - glottal source, with mechanical oscillations of the folds, caused by the flowing air from the respiratory system - the lungs. With folds opening and closing, the air flow is modulated across the glottal opening, thus producing time varying acoustical signals.

Many mammal species produce almost periodical signals in the larynx with basic frequency and multiple harmonics, which can be relatively easy distinguished by Fourier transform or other speech analysis techniques in frequency domain. On the other hand, narrower, but non-oscillating larynx produces turbulent noise. Thus, the most appropriate way of description of the produced animal vocalization would be by the general discrete-time model for voice (speech) production.

2.1 Formants and Vocal tract length

According to acoustical theory central frequencies of the formats are related and depend on the anatomical measurements of the vocal tract, on the length and on the shape (particular variations in the cross sections). Moreover, the length is the most important anatomical measure which influences the formant frequencies [3] (Figure 1). The resonant frequencies of the uniform air tube with non-varying sections can be used as the first approximation of the formats given by:

$$F_i = \frac{(2i+1)\cdot c}{4\cdot VTL}$$

With the both ends closed:

$$F_i = \frac{i\cdot c}{2\cdot VTL}$$

Where i=1 is the number of the format, c is the speed of the sound (350 m/s), VTL - vocal tract length (in meters) and- F_i the frequency of *i-th* format. Despite the difference in the equitation regarding the status of the ends (opened or closed), it is obvious that the distance between two consecutive formants is constant and is given by:

$$F_i - F_{i-1} = \frac{c}{2\cdot VTL}$$

The distance between two formants is constant and directly depends on the vocal tract length. Larger value for VTL

gives lower formant frequencies. It can be assumed that different breeds of dogs have different or dissimilar anatomical features (VTL), produces different vocalization pattern, regarding formant frequencies. This apply on individual subjects as well, each animal has different anatomical characteristics related to age, sex, size and other factors.

Therefore it will be possible to classify the vocalization sequences in certain categories. Such classification system will be able to recognize the species, breed, and individual animals (in these case dogs) by their vocalization or to estimate the possibility of presence of anatomical deviation in the vocal tract length in animals of the same breed [4].

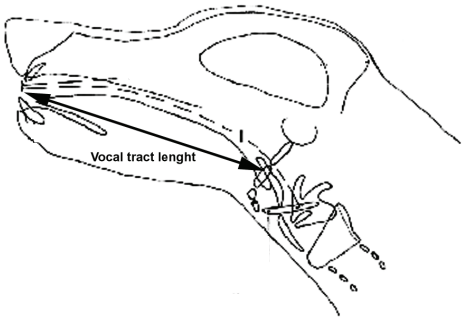

Figure 1: *Vocal tract length measurement*

In order to achieve accurate classification of the dog vocalization, well-known techniques for human speech recognition may be used [1]. This paper presents a method for classification of dog vocalization sequences based on Time Encoded Signals, for signal representation, and Artificial Neural Networks for their classification.

2.2 Proposed system

The overall system architecture for domestic dog vocalization classification for speaker recognition is shown on Figure 2. The system consists of blocks for acquisition and preprocessing of signals, A/D conversion block, TES coder where digitized signals are coded into time domain with array of symbols. Block where S-matrix with fixed length is created and represents the input vector of the neural network classificatory. This system could be implemented as a part of low-end DSP hardware system, as well.

2.3 Recording, digitalization and preprocessing

Vocalization sequences of the subjects in a state of excitement were recorded with ordinary microphone. Figure 3 presents characteristic spectrogram of the barking sequence of a German Sheppard dog breed. Voice signal was digitized with 20.05 KHz sample rate and 16 bit resolution, providing quality in processing and more discriminate classification.

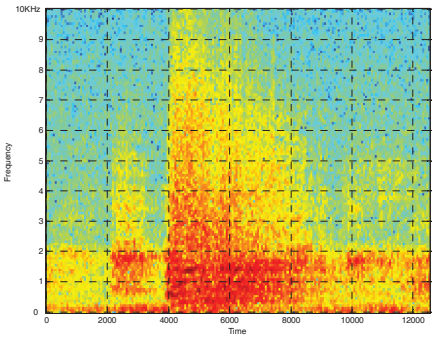

Figure 3: *Spectrogram of "German Sheppard" vocalization*

2.4 TES - Time Encoded Speech

The choice of the techniques for animal voice signals recognition is most important for designing a reliable system. There are several methods that could be applied to solve the stated problem. One possible solution is to use vector classification of the PSD (Power Spectrum Density) function. To consider the time varying nature of the vocal signal, DTW (Dynamic Time Warping) technique could be used on arrays of MFCC (Mel Frequency Cepstral Coefficients) [4]. But, despite the algorithm simplicity, this method could not provide efficient and reliable system. Other possible solution is to use HMM (Hidden Markov Model) for representing vocal sequences with more complex training process. Regarding the nature of the application and the known characteristics of voice signals, this system uses Time Encoded Signal Processing and Recognition [5] for signal parameterization and Neural Network for classification. This approach provides low complexity and computing requirements that will be well suited for robust and embedded hardware applications.

TES coding is based on precise mathematical description of waveforms, involving the polynomial theory that shows how band limited signals (such as animal vocalization) may be completely described by the locations of their real and complex zeros. The real and complex zero descriptors of TES and the time-bandwidth data produced by a Fourier transform are mathematically equivalent and both produces equal number of digital sample points. This technique is well known and it has been used for a long time in speech coding in telecommunication systems [5]. The interval between two adjacent zero-crossings is called an epoch and, for every epoch, three parameters are derived: duration of the epoch, shape (S) and the magnitude (M) of the signal. D is the number of samples and S is the number of positive or negative local maxima and minima, and M is the largest value of samples in the given period.

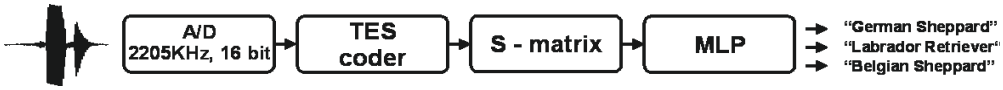

Figure 2: *Simulation system architecture*

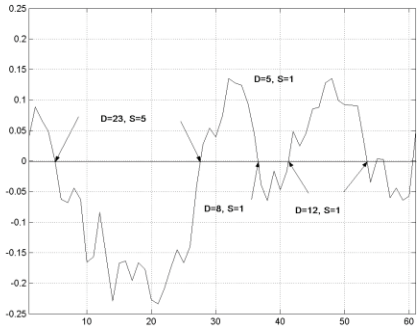

Figure 4: *Part of vocalization waveform divided in TES epochs*

These parameters are encoded with assigning a unique symbol for certain combination of the epoch duration (D) and its shape (S) (Figure 4). Thus the signal is transformed into time encoded stream of discrete numerical descriptors – TES symbols. The statistical analysis showed that number of samples between zero crossings lies between 1 and 37 for speech signals sampled with 20.05 KHz [5] and that the duration of an epoch above 13 samples is very unlikely to occur.

Using vector quantization and K-means algorithm, generalized code-book was created – TES alphabet. Standard symbol alphabet consists of 28 different symbols, and it has been proved to be quite sufficient for the representation of speech and other band limited signals. The symbols table is established according to statistical analysis and the likely occurrences of the (D, S) pairs for given typical signals. In case of pair that could not be found in the symbol table, it could be linked to the symbol that best represents the epoch for the given pair. These strings of numerical descriptors can be easily converted to TES matrices with fixed dimension. A histogram of the signal array with 28 possible symbolic descriptors can be produced, forming so called S-matrix with fixed dimension 1x28 (Figure 5), which carry information about symbols frequency.

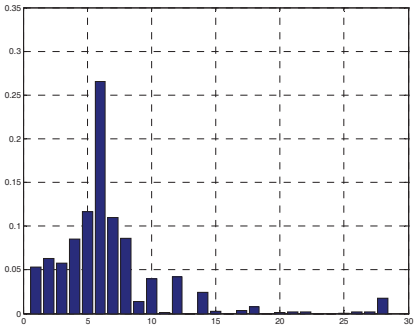

Figure 5: *S-matrix of "German Sheppard" vocalization*

TES coder transforms the original signal in array of symbolical descriptors without losing the quality. It can be easily reversed back into analog signal. Because of their fixed length, S-matrices are ideal form of input vector for artificial neural network classificatory.

The biggest advantage of the S-matrices compared to other descriptors in the frequency domain (eg. MFCC) is that regardless of the signal length they have fixed dimensions. It allows easy template creation for the system training phase, and allows use of wide range of different classification procedures.

2.5 Classification with Multi Layer Perceptron

Because of their fixed length, S-matrices are ideal form of input vector for neural network classificatory. This system uses Multi Layer Perceptron (MLP) with 3 layers, where the number of input layer nodes is equal with the length of the S-matrix – 28 and the number of outputs by the number of different classes, in this case 11 individual animals. The hidden layer contains 30 neurons with logsig transfer function and the output layer 11 neurons with also logsig function. The choice of the number of neurons in the hidden layer is made empirically comparing the neural network performance.

The choice of transfer function in the hidden layer was made according to the characteristics of the S - matrices whose elements have only positive values in the range between [0, 1]. This causes output neurons to produce values also in this range but represents probabilities of recognition for given input in output classes, in this case – that is individual animal (dog).

In case where the number of the neurons in the hidden layer is too small, the network is not trained to classify input samples appropriately (under fitting). Otherwise, increasing the number of neurons in hidden layer may introduce an effect of saturation (over fitting) and the neural network is well trained to classify or recognize only the training set of input samples. The training set is composed of input – output pairs of S-matrices and outputs values 0 and 1 on the corresponding node. For the training of the ANN, Resilient Back propagation was used for determining the direction of change of weight functions, but not the weight values. They were determined by a special variable which is increased by a factor delta_inc = 1.2 whenever that value has the same sign for two consecutive iterations and is reduced by a factor delt_dec = 0.5 whenever the derivative with respect to weight values sign changes to the previous iteration

Whenever an oscillation in the direction of change of weight function values appears, the variable decreases, and if for some period its sign does not change, the variable value steadily increases [6]. The process of the training is stopped in the moment when the target MSE is achieved or maximum number of training epochs are reached.

3 SYSTEM'S PERFORMANCE ESTIMATION

Barking samples of eleven individual police dogs were used for system performance estimation: three different subjects of the breed "Belgian Sheppard - Malinoi" with 47 sequences (Robi, Niki and Gor), five subjects of "German Sheppard" with 88 sequences (Go1, Amor, Dog, Gando, and

Rex) and three subjects of "Labrador Retriever" (LB1, Ago and Ari) with 25 sequences. That gives total number of 160 barking samples of 11 animals of 3 different dog breeds. Rough segmentation of minimum of 10 samples per each individual dog was taken and training set (Figure 6) was produced with 110 vocalization sequences. The system is implemented using the MATLAB software package.

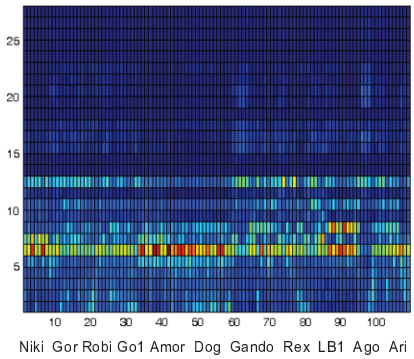

Figure 6: *Training set contained total of 110 sequences*

The system was tested for individual dog recognition by their vocalization. The test set consists of 162 samples, the neural networks topology consists of 28 input neurons, 30 hidden neurons and 11 output neurons, target mean square error was set to 1e-6, and the ANN was trained for 200 epochs.

	Niki	Gor	Robi	Go1	Amor	Dog	Gando	Rex	LB1	Ago	Ari
Niki	8	0	1	1	0	0	0	0	0	0	0
Gor	0	3	2	0	0	1	1	0	0	0	0
Robi	0	0	29	1	0	0	0	0	0	0	0
Go1	0	0	0	10	1	0	1	0	0	1	0
Amor	0	0	1	0	24	0	0	0	0	0	0
Dog	0	0	1	0	0	7	1	3	0	0	0
Gando	0	0	2	0	0	0	19	1	2	2	0
Rex	0	0	0	2	0	0	0	12	0	0	0
LB1	0	0	0	0	0	0	0	0	12	0	0
Ago	0	0	1	0	0	0	0	0	0	5	0
Ari	1	0	0	0	0	0	1	1	0	0	4

Table 1: *Resulting confusion matrix for the individual animal recognition*

The results of the performance estimation of 162 samples recognition are:

Training set: 94.68 % with: 110 samples
Test set: 80.25 % with: 162 samples

During training process, it is noticeable that after 100 epochs - the MSE decreases slowly, so higher value for MSE or shorter time for training could be set. This will produce better generalization of the neural network and increase the reliability of the system for samples that were not included in the training phase.

4 CONCLUSIONS

The proposed system uses Time Encoded Signals (TES) and Artificial Neural Networks (ANN) for classification of individual animals. It was tested and evaluated with vocalization samples of 11 individual subjects (three different police dog breeds, "Belgian Sheppard - Malinoi" "German Sheppard" and "Labrador Retriever"). It is shown that this system is able to successfully recognize subjects with different anatomical characteristics (different individuals) of the vocal tract by their vocalization samples. Such a system could be successfully used in various on-field applications implemented on low complexity smart sensor DSP hardware: monitoring and research on animal behavior of different species, their communications, automatic species detection, small animal practice in veterinary medicine, monitor individual animals behavior on limited space (e.g. Zoo and National parks) and others.

References

[1] J. G. Harris, M. D. Skowronski, "Automatic Speech Processing Methods For Bioacoustics Signal Analysis: A Case Study Of Cross-Disciplinary Acoustic Research" IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Toulouse, France, vol. 5, pp. 793-796, May 15-19, 2006

[2] Phipps, T.C., King, R.A. "A Low-Power, Low-Complexity, Low-Cost TESPAR-based Architecture for the Real-time Classification of Speech and other Band-limited Signals" ICSPAT at DSP World, Dallas, Texas, October 2000

[3] T. Riede, T. Fitch: Vocal Tract Length And Acoustics Of Vocalization In The Domestic Dog (Canis Familiaris), The J .Of Exp. Biology 202, 2859–2867 (1999)

[4] Kraljevski I., Mihajlov D., Arsenovski S., „Determination of Farm Animals Condition by Their Vocalization", Proceedings of the 24th ITI, A-7, Cavtat, Croatia, June 24-27, 2002.

[5] J. Holbeche, R.D. Hughes and R.A. King: Time Encoded Speech (TES) descriptors as a symbol feature set for voice recognition systems, IEEE Int. Conf. Speech Input/Output; Tech. and App., pp. 310-315, March 1986.

[6] Matlab Product Documentation