Localization in Static and Dynamic Hearing Scenarios: Utilization of Machine Learning and Binaural Auditory Model

Ekaterina KOSHKINA, Jaroslav BOUSE

Dept. of Radioelectronics, Czech Technical University in Prague,

Technická 2, 166 27 Praha, Czech Republic

koshkek1@fel.cvut.cz, bousejar@fel.cvut.cz

Abstract. *Hearing with both ears, in other words binaural hearing, allows human to localize sound sources in a space. Models of binaural hearing often simulate functions of lateral and medial superior olives (LSO and MSO), but their outputs cannot be in most cases directly mapped to certain azimuths in space.*

In this paper we present an azimuth classification algorithm, which utilizes both binaural models (LSO and MSO) for preprocessing of a sound signal. From their outputs features are extracted for machine learning algorithms: K-Nearest Neighbors and Artificial Neural Network. The algorithm is trained and tested on speech samples from ITU-T Rec. P501 and NOIZEUS corpora. The success of the K-NN and ANN classifiers is discussed. Both machine learning algorithms give a similar classification error in static and dynamic hearing scenarios. The error is comparable to human psychoacoustical data.

Keywords

Binaural auditory model, LSO, MSO, azimuth, static sound source localization, dynamic sound source localization, feature extraction, RMS, classification, K-NN, ANN, MATLAB

1. Introduction

Sound localization as a process of determining the sound source position in a space is an important ability of human hearing. Humans utilize both ears while localizing the sound source, the term binaural hearing is used in reference to this ability.

Binaural hearing on the horizontal plane is primarily associated with two binaural cues: the time difference of the signal coming to the right and left ear called the interaural time difference (ITD), and the level difference of the signal coming to the ears called the interaural level difference (ILD) [1].

Information from the right and left ears is processed in the human brain stem in the superior olivary complex (SOC). In the SOC two nuclei successfully decode binaural information: the lateral superior olive (LSO) and the medial superior olive (MSO). While LSO decodes intensity differences, MSO successfully decodes the time differences in the incoming binaural signal [2].

Both ITD and ILD change according to the position and spectral characteristics of the sound source. This relationship is unique for each person and is fully described by his head-related transfer function (HRTF) [1].

Based on the so-called Duplex theory [3] ITD is used in the localization of signals with frequencies under approximately 1.5 kHz, and ILD is applied at the localization of signals with higher frequencies. Although there exists evidence that ILD is also utilized in the localization at low frequencies and ITD can be detected from envelopes of high frequencies [2].

In our previous paper [4], we investigated the possible use of a K-Nearest Neighbors (K-NN) classifier as a cognitive block in the binaural auditory model for horizontal plane localization. The binaural model was composed of medial and lateral superior olives models [5,6,7], which outputs were used as features for the K-NN classifier. The localization predictions based on features from MSO and LSO were analyzed separately.

In this paper we propose a method which utilizes both MSO and LSO model features to localize sound source on the frontal horizontal plane. We examine the possible use of the K-NN and Artificial Neural Network (ANN) classifiers as cognitive devices for the localization task. The performance of both cognitive devices is evaluated on the HRTF positioned sound stimuli from ITU-T Rec. P.501 and NOIZEUS speech corpora. The classification error for both static and dynamic (moving) sound sources is provided. The results imply better performance in front of the artificial listener than on the lateral sides, which agrees with psychoacoustical data [1].

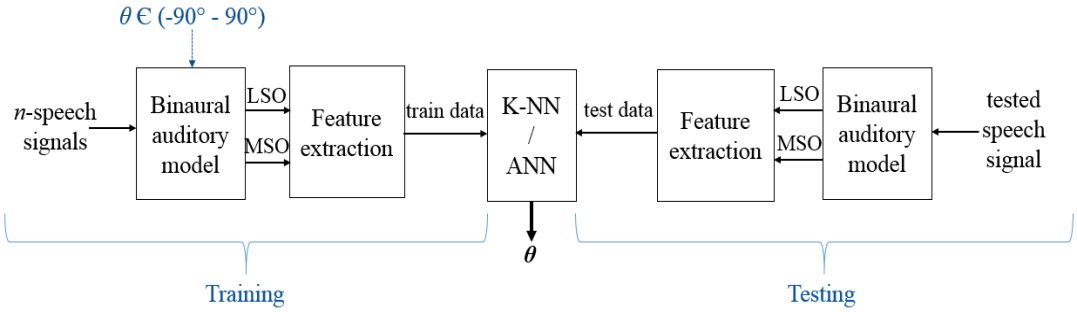

Fig. 1. The block diagram of the implemented algorithm.

2. Classification Algorithm

The structure of the proposed algorithm is depicted in Fig. 1. The binaural signal is first preprocessed by binaural auditory models (LSO and MSO). The output of these models corresponds to perceived lateralization function i.e. localization within the listener's head. There are multiple lateralization functions each for a specific band with a central frequency f_c. Root-Mean Square (RMS) is calculated in each band and is used as the feature for the classification.

We use two supervised machine learning approaches for the classification: K-Nearest Neighbors (K-NN) and Artificial Neural Network (ANN). Both approaches are used separately.

The K-NN classifier works by measuring distances between the training set and the unknown test signal. The test signal's class is determined as the class of the most common element in the group of the K measured minimal distances [8]. More detailed description is in the previous paper [4].

The ANN algorithm [9] somehow mimics the activity of the human brain and the neurons' function. ANN typically has multiple layers and the signal is propagated from the front of the network to the back. The connections of the neurons from the different layers are characterized by weights. A common supervised learning principle is the backwards error propagation of weight adjustments. At first the input training set is forward propagated through the network. Then the output of the network is compared to the target value and the error is calculated. After that the weights are modified for the error reduction due to the backward propagation of the error. The learning process is finished when the error is minimal. ANN signal classification is started when the weights are adjusted.

The predicted azimuth of a sound source (θ) is obtained by combining classification results from the LSO model (θ_{LSO}) and from the MSO model (θ_{MSO}):

$$\theta = \begin{cases} \dfrac{\theta_{LSO} + \theta_{MSO}}{2}, & |\theta_{LSO} - \theta_{MSO}| \le 20 \\ \theta_{LSO}, & |\theta_{LSO} - \theta_{MSO}| > 20 . \end{cases} \quad (1)$$

The limit of 20 degrees is based on the results of our previous experiment [4], where the LSO model provided more reliable classification results.

3. Methods

Sound localization as proposed in this paper is done for the static sound source and for the dynamic sound source.

The classification classes are the azimuths of an incoming sound $\theta \in$ (-90° – 90°) with steps equal to 5 degrees, whereas -90° corresponds to a signal next to the left ear, 0° to a signal in front of the listener and 90° next to the right ear.

3.1 Stimuli

For experimental purposes speech corpora from the NOIZEUS [10] and ITU-T Rec. P.501 [11] are used. From these corpora 64 one-second long speech signals with the sampling frequency 44.1 kHz are generated. According to the recommendation [12] the generated database is split into training and testing sets in the ratio 2/3 and 1/3, respectively.

3.2 Training

The training set generation is the same for the static and dynamic sound sources.

The speech signal from the training set is filtered by the head-related transfer function (HRTF) from the ARI database [13]. The HRTF corresponds to certain azimuths in the range $\theta \in$ (-90° – 90°). The filtered signal is then processed by binaural auditory models (LSO and MSO). The RMS features are extracted from the acquired signal. The obtained training set is composed of 48

patterns for each azimuth and is created individually for the LSO and MSO models.

3.3 Testing

A. Experiment I - Static Sound Source

The testing set is generated in a similar way as the training set.

In the K-NN classification the values of the received features from the testing set are compared with the values of the features from the training set. Based on the experiment the classifier's parameter K is set to K=3.

ANN uses the same training and testing sets, as K-NN. It is important to choose the appropriate network architecture. The network in this paper consists of 27 input neurons and 37 output neurons and it has two hidden layers with 30 and 33 neurons, respectively. The network architecture is defined based on the so-called geometric pyramid rule [14].

B. Experiment II - Dynamic Sound Source

The testing signal is linearly moved signal from -90° to 90°. This signal is generated with the help of the binaural auditory model and HRTF function too. This moving signal is a 37-second long signal, where each second of the signal corresponds to a different azimuth from the range.

In the classification, the moving signal is uniformly segmented on 37 segments which correspond to the number of the classes of the azimuths. The segment length is set as the length of the training signal, which equals one second. Every segment is processed separately. This means that the classification of each segment follows the scheme of the static sound source localization.

4. Results

The results of both experiments are depicted in Fig. 2 - Fig. 4. The mean values and standard deviations are calculated from all stimuli within the test group in all cases.

A. Experiment I - Static Sound Source

The outputs of the static sound source localization utilizing the K-NN classifier and ANN are illustrated in Fig. 2 and Fig. 3, respectively.

The red points show the classification results. In the optimal case, all these points should lie on the blue line.

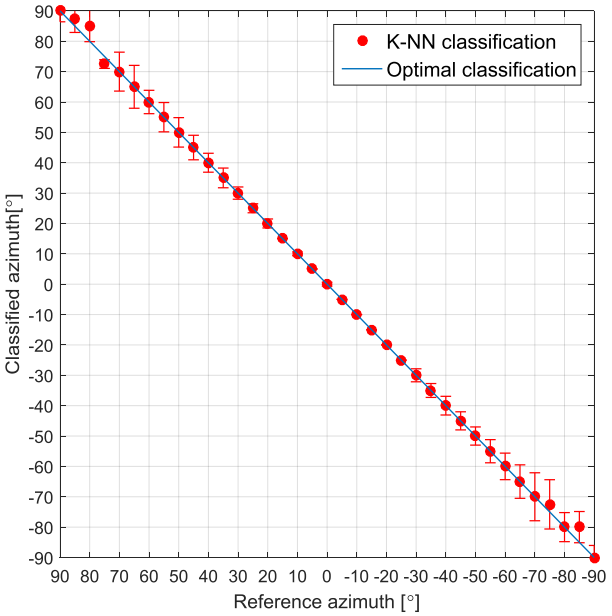

Fig. 2. The classification error dependency on the reference azimuth utilizing the K-NN classifier with K=3 for the static sound source.

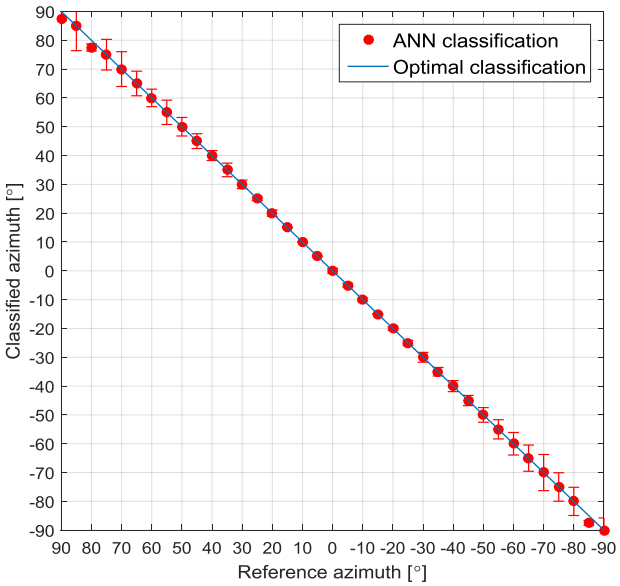

Fig. 3. The classification error dependency on the reference azimuth utilizing ANN with two hidden layers for the static sound source.

B. Experiment II – Dynamic Sound Source

The graphs for the dynamic sound source localization are depicted in Fig. 4 and Fig. 5.

In this case the moving signals are classified. These generated moving signals have the linear time-varying azimuth from the range. Each one-second segment corresponds to one azimuth.

Optimally the red points, as the classification results, should be on the blue line too.

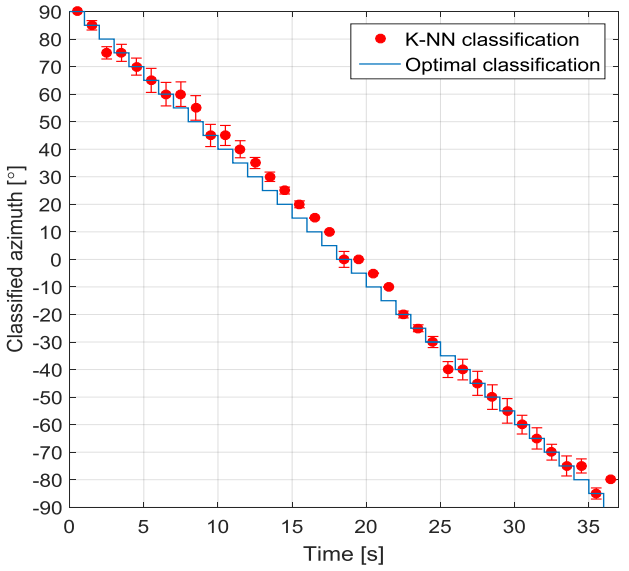

Fig. 4. The classification error dependency on the reference azimuth utilizing the K-NN classifier with *K*=3 for the dynamic sound source.

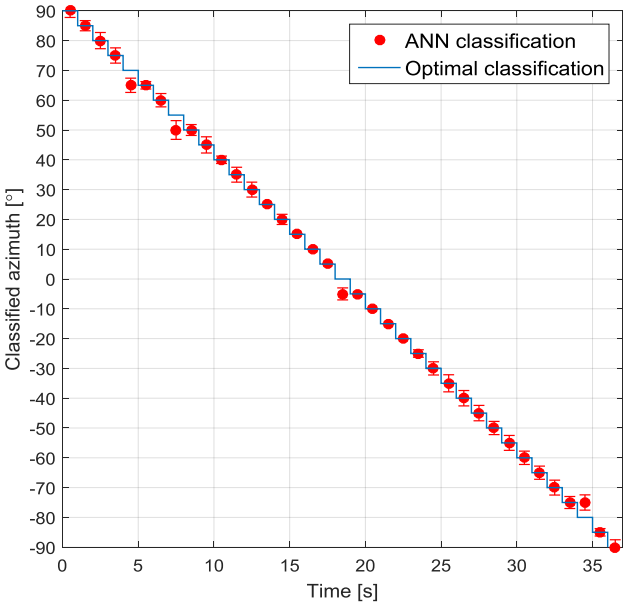

Fig. 5. The classification error dependency on the reference azimuth utilizing ANN with two hidden layers for the dynamic sound source.

5. Conclusion and Future Work

This paper presents the static and dynamic sound source localization algorithm utilizing the K-NN and ANN classifiers. RMS values of binaural auditory models (LSO and MSO) are used as features. The output of the implemented algorithm is the classification error dependency on a reference azimuth.

In comparison K-NN is more computationally demanding than ANN. The classification accuracy is similar for both classifiers.

The average azimuthal error across testing set for both K-NN and ANN can be observed in Tab.1.

Error / Classifier	0°	5°	10°	>10°
K-NN	55 %	38 %	7 %	0 %
ANN	65 %	32 %	3 %	0 %

Tab. 1 The average azimuthal error across the testing set for the K-NN and ANN classifiers.

Generally, the classification is more successful in the detection of azimuths around the zero angle and decays reaching the higher azimuths, which corresponds to the human psychoacoustical data [1].

An obvious choice for future work is to optimize the algorithm for the localization of the different common and uncommon sounds. Further improvement may involve the sound localization on the vertical plane.

Acknowledgements

Research described in the paper was supervised by Ing. Frantisek Rund Ph.D., FEE CTU in Prague and supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/190/OHK3/3T/13

References

[1] Blauert, J. and J. S. Allen (1997). "Spatial Hearing - The Psychophysics of Human Sound Localization*". Rev. Cambridge: MIT Press.* ISBN 978-0-262-02413.

[2] Meddis, R. (2010). "Computational Models of the Auditory System". *Springer Handbook of Auditory Research, 35. Springer US.* ISBN: 9781441959348.

[3] Rayleigh, L. (1907). "On our perception of sound direction". In: *Philosophical Magazine* 13, p. 232.

[4] Koshkina, E. and Bouse, J. (2016). "Lazy Learning Sound Localization Algorithm Utilizing Binaural Auditory Model". In: Proc. of 20th International Student Conference on Electrical Engineering POSTER 2016. Prague: Czech Technical University in Prague. 4 pp.

[5] Vencovsky, V. and J. Bouse (2011). "Binaural Processing Model Simulating the Lateral Position of Tones with Interaural Time Differences". In: *Proc. of 15th International Student Conference on Electrical Engineering POSTER 2011*. Prague: Czech Technical University in Prague. 5 pp.

[6] Bouse, J. (2013). "A Model of Directional Hearing" *MA thesis. CTU in Prague, Faculty of Electrical Engineering (Advisors: Rund, F. and Vencovsky, V.)* 47 pp.

[7] Bouse J. and V. Vencovsky (2015). "Two-channel models of medial and lateral superior olive based on psychoacoustics". In: *BMC Neuroscience* 16. Suppl 1, P276. ISSN: 1471-2202. url: http://www.biomedcentral.com/1471-2202/16/S1/P276.

[8] Larose D. T. (2005). "Discovering Knowledge in Data: An Introduction to Data Mining". *John Wiley & Sons, Inc.*, ISBN 9780471666578.

[9] Rojas R. (1996). „Neural Networks - A Systematic Introduction ", Springer-Verlag, Berlin, New-York, ISBN-10: 3540605053.

[10] Y Hu. (2007). PC Loizou, Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49(7-8), 588–601.

[11] ITU. (2012). Test signals for use in telephonometry, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.501.

[12] Alpaydın E. (2010). „Introduction to Machine Learning “, The MIT Press Cambridge, Massachusetts London, England, Second Edition, ISBN 978-0-262-01243-0.

[13] The Acoustics Research Institute of the Austrian Academy of Sciences. The ARI HRTF database.
url: http://www.kfs.oeaw.ac.at/hrtf

[14] Masters T. (1993). “Practical Neural Network Recipes in C++”. *Academic Press. Inc. San Diego,* ISBN:0-12-479040-2.

About Authors...

Ekaterina KOSHKINA was born in Moscow, Russian Federation in 1993. In 2015 she received her bachelor diploma from Communication, Multimedia and Electronics program at Faculty of Electrical Engineering (FEE) Czech Technical University (CTU) in Prague. Her bachelor thesis covered the subject of archive audio record content identification. She is currently working on her master’s degree in Communication, Multimedia and Electronics program at the same faculty.

Jaroslav BOUSE was born in Prague, Czech Republic in 1988. In 2010 he received his bachelor diploma from Electronics and Telecommunications program at Faculty of Electrical Engineering (FEE) Czech Technical University (CTU) in Prague. He graduated in the premium advanced form of the Communications, Multimedia and Electronics master’s degree program (a one of 7 students enrolled from total number of 144 students enrolled in the program), focusing on Multimedia Technology, at the same faculty in 2013. He is now Ph.D. student at the department of radioelectronics at the same faculty. His research topic is Audio signal processing from the psychoacoustic point of view, with the specialization in spatial hearing experiments and models.