Lazy learning sound localization algorithm utilizing 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. *This paper introduces an algorithm for sound localization in the horizontal plane (azimuth) in utilizing the binaural auditory model.*

The decision device is based on segmentation, feature extraction and classification, which is implemented in MATLAB environment. The algorithm uses k-nearest neighbors (KNN) classifier with k=20. Exploited features are root mean square and signal decimation. The output of this algorithm is the relative distribution of KNN across groups of azimuths.

Keywords

Binaural auditory model, LSO, MSO, azimuth, localization, segmentation, feature extraction, RMS, decimation, classification, KNN, MATLAB

1. Introduction

Sound localization, the process of determining location of a sound source is a significant feature of human hearing. Localization can be described in terms of three-dimensional position: the angle in the horizontal plane (azimuth), the angle in the vertical plane (elevation), and the distance (for static sound sources) or velocity (for moving sound sources) [1]. For the purpose of this paper only horizontal plane is considered.

In the horizontal plane is human hearing system able to decode spatial position of sound source based on two binaural cues: the difference of arrival of the sound waveform at the left and right eardrums called interaural time difference (ITD), and the difference in sound pressure at them called interaural level difference (ILD) [1]. The relationship between ITD and ILD dependent on direction of incidence of sound source is described in head related transfer function (HRTF) and is unique for each human [1]. The binaural cues are decoded in human brain stem in the superior olivary complex (SOC), more precisely in two SOC's nuclei: medial and lateral superior olive (MSO and LSO) [2]. The ITDs are processed in the MSO and the ILDs in the LSO [2]. Such partitioning hardly follows the so-called duplex theory [3], which states that the tones with frequencies under approximately 1.5 kHz are localized based on ITDs, the ones with higher frequencies based on ILDs.

Properties concerning localization aspects of binaural auditory model can be evaluated by usage of standard methods of feature extraction and classification. One of the simplest classification methods is k-nearest neighbors classifier (KNN) [4]. This supervised learning algorithm measures the distance between features of the unknown object and the features of labeled objects from training set. Based on measured distance unknown object's class is determined as the class of the most common element in the group of the k-nearest neighbors. KNN can show good results for some recognition problems despite its simplicity. One of the main drawbacks of this classifier is its computational efficiency, which decreases with the number of records in the dataset [5, 6].

In this paper we propose a lazy learning sound localization algorithm utilizing binaural auditory model. The incoming binaural sound signal is first preprocessed by binaural auditory models developed at our department [7, 8, 9]. These models mimic functionally the behavior of medial and lateral superior olives. From the binaural models' outputs are extracted features (root mean square, decimation) and analyzed by k-nearest neighbors classifier (KNN).

At the output of this algorithm is the relative distribution of KNN across groups of azimuths. According to initial experiment the algorithm works well for signal position by a small azimuthal angle.

2. Binaural hearing model

Before the localization of the incoming acoustic signal some preprocessing is required.

Firstly, the incoming signal is processed by the models of human ears, which are each approximated by a cascade of 512th order finite impulse response filters simulating the outer and middle ear; followed by a bank of

27 dual resonance nonlinear filters (bandwidth equal to 1 equivalent rectangular bandwidth) simulating the frequency selectivity of cochlea, and finally half-wave rectification followed by a low-pass filter simulating inner hair cells [10].

The signal from both ears is then analyzed in models of medial and lateral superior olive [9]. The outputs of these models correspond to subjective lateralization (i.e. localization within listener head) and are used further for estimating the horizontal angle of incidence of sound by the feature extraction and classification.

3. Feature extraction

Feature extraction is the process of computing numerical representation that can be used to characterize different signals. The implemented algorithm uses two features: root mean square (RMS) and the signal decimation.

3.1 Root Mean Square (RMS)

Root mean square is a common simple feature in the signal analysis. RMS is defined as the square root of the arithmetic mean of the squares of the values of the signal. RMS of the signal $s(n)$ with length N is represented as following [11]:

$$RMS = \sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} s(n)^2} \ . \tag{1}$$

3.2 Signal decimation

Decimation is a form of signal filtration and it is the process of decreasing the sampling rate of a signal. It is realized by the decimation factor which divides the sampling rate. This factor is a positive integer scalar [12]. Appropriately chosen decimation factor can decimate signal in way, that the result represents signal envelope.

4. Classification

Classification is process of sorting signals into predefined classes. In this paper are these classes azimuths of an incoming sound $\theta \in$ (-90° – 0°) and $\theta \in$ (0° – 90°) with step equal to 5 degrees, whereas -90° corresponds to signal next to left ear, 0° to signal in front of listener and 90° next to right ear.

As a classifier k-nearest neighbors (KNN) is used for azimuth recognition purposes. It is the statistical model with the supervised learning, where the function is inferred from labeled training data set. The training set contains labeled data corresponding to mentioned classes. Data from the training set is a part of n-dimensional space. Tested element must lie in the same n-dimensional space, so that distances between test element and elements from training set could be calculated. After determining all of the distances, k-nearest neighbors are chose. The tested element is then assigned with a class which is the most abundant in the k-nearest neighbors [5, 6].

Euclidian distance is used to calculate difference between elements. It is defined by following equation [13]:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_i (x_i - y_i)^2} \ , \tag{2}$$

where $\vec{x} = (x_1, x_2, ...x_n)$ and $\vec{y} = (y_1, y_2, ..y_n)$ are vectors in Euclidian n-dimensional space.

In the implemented algorithm the classification is done for the lateral superior olive (LSO) model and for the medial superior olive (MSO) model separately. The statistical pattern recognition classifier is trained using the generated noise. The noise is filtered by the head related transfer function (HRTF) corresponding to certain azimuth in the range $\theta \in$ (-90° – 90°). HRTF comes from an ARI database and was measured on an artificial head and torso simulator [14]. Afterwards, specific features (RMS and decimation) of the generated signals are calculated. This training set has 20 patterns for each azimuth and is created individually for the both models (LSO and MSO).

The next step is classifier testing. At first the testing sets are generated the similar way as the training set, but individually for each of the azimuths from the range. More patterns are created for better evaluation of the implemented classifier. Computation of RMS requires segmentation of the analyzed signal, because its characteristics can vary in time. Signal is segmented to 512 samples long segments and in each of these RMS is found. For case of the signal decimation the analyzed signal is decimated by a factor of 500. Values of received features from the testing signal are compared with the features from the training set using the KNN classifier with k=20, and then the class is determined. The relative distribution of KNN across groups of azimuths is calculated as the amount of neighbors found by the classifier for each given azimuth divided by the absolute value of k.

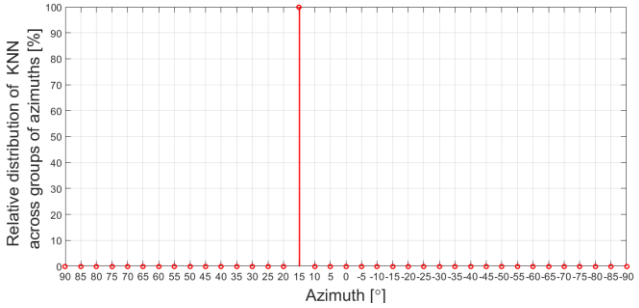
5. Results

The proposed algorithm has the graphical output for each tested azimuth. Plots show the relative distribution of KNN across groups of azimuths, which describe how implemented algorithm recognized tested signal.

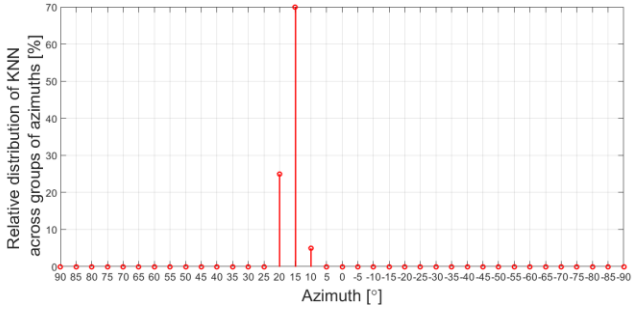
The results of classification are better with the usage of the signal decimation, than with usage of the RMS. Classification with RMS often recognizes azimuth of the tested signal as a nearby azimuth but not the correct one.

See Fig. 3 and Fig.4 for the results of testing signal corresponding to $\theta \in 15°$ using both features individually for LSO and MSO, respectively.

The recognition success rate of the algorithm for both features and both models is illustrated in Fig. 5 and Fig. 6. It is calculated using 100 samples big testing sets for LSO and MSO model.

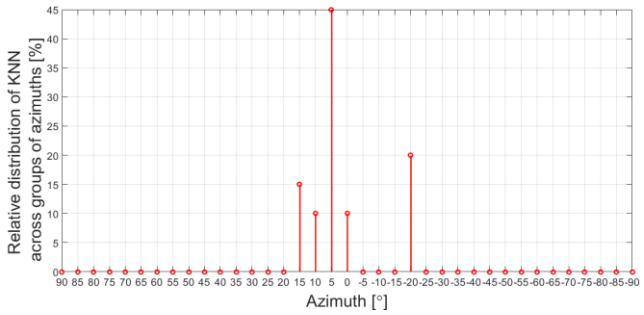


(a)

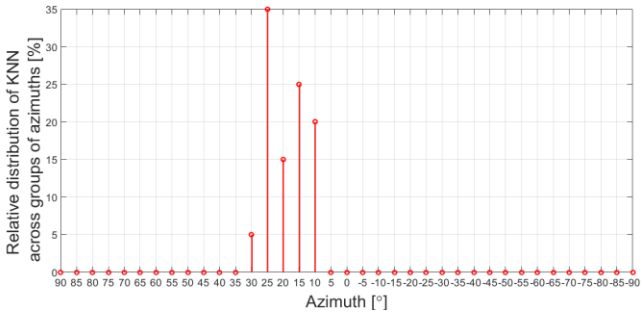


(b)

Fig. 3. The relative distribution of KNN across groups of azimuths for $\theta = 15°$ with using the signal decimation. *(a)* LSO model, *(b)* MSO model.

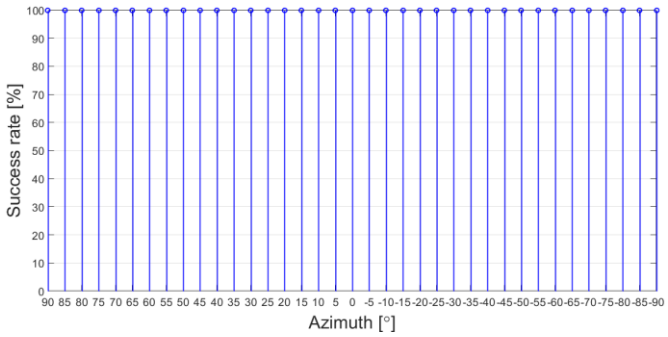


(a)

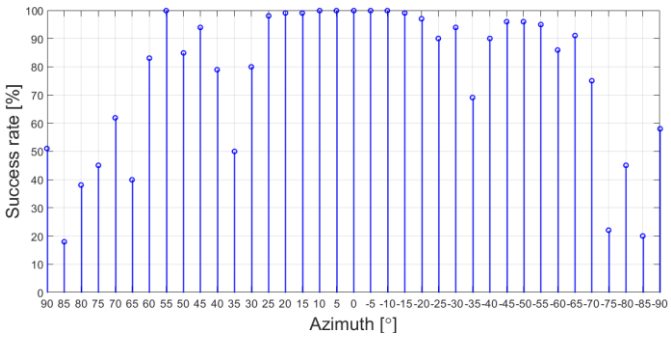


(b)

Fig. 4. The relative distribution of KNN across groups of azimuths for $\theta = 15°$ with using RMS. *(a)* LSO model, *(b)* MSO model.

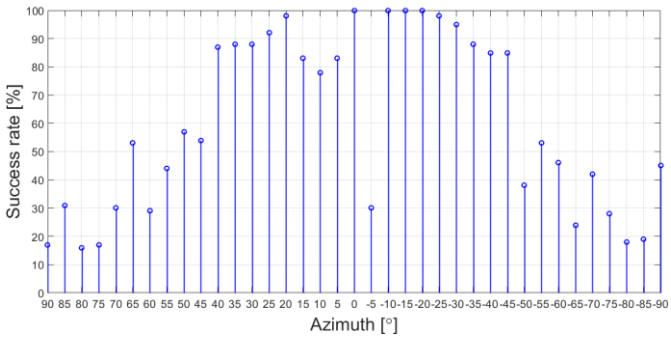


(a)

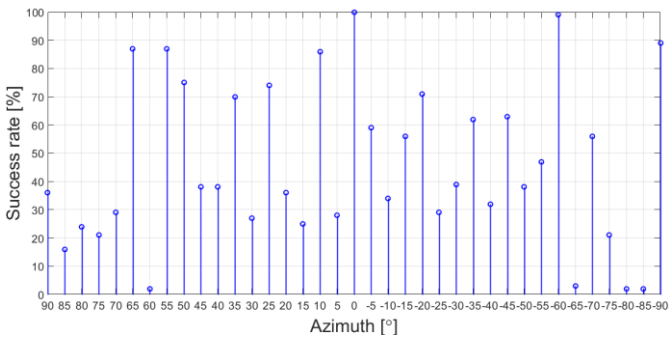


(b)

Fig. 5. Algorithm recognition success rate using the signal decimation. *(a)* LSO model, *(b)* MSO model.



(a)



(b)

Fig. 6. Algorithm recognition success rate using RMS. *(a)* LSO model, *(b)* MSO model.

6. Conclusion and future work

This paper presents azimuth classification for LSO and MSO model, using KNN classifier. Feature extraction

for classifier is based either on calculation of RMS or signal decimation.

Judging on the results, the classification using the signal decimation is more accurate than the classification using RMS. Generally, the classification is more precise for LSO model. The classifier is more successful in the detection of azimuth around the zero angles. In the higher values of azimuths, the outputs of MSO and LSO are getting into the saturation. Hence, it is harder to classify location accurately, using these outputs [8].

Tab. 1 shows the average recognition success rate for all possible settings of implemented algorithm.

Feature / Model	Decimation	RMS
LSO	100 %	60 %
MSO	73,5 %	46 %

Tab. 1. The average recognition success rate comparison.

An obvious choice for future research is to try solving the situation, when the classifier hasn't recognized the correct azimuth and to explore reasons for those errors. Further upgrade may involve using other features or maybe their combination. Using more advanced classifier could provide an effective solution of the problem.

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.SGS14/204/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] Guo G., Wang H., Bell D., Bi Y., Greer K. (2003)." KNN Model-Based Approach in Classification". Springer Berlin Heidelberg. pp. 986-996.

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

[6] Devroye L., Györfi L., Lugosi G. (1996). "A probabilistic theory of pattern recognition*". Springer-Verlag New York Berlin Heidelberg*. ISBN 0-387-94618-7.

[7] 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.

[8] 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.

[9] 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.

[10] Poveda, E. A. L. and R. Meddis (2001). "A human nonlinear cochlear filterbank". In: *The Journal of the Acoustical Society of America* 110(6), pp. 3107-18.

[11] A Dictionary of Physics 6 ed., (2009*). Oxford University Press*. ISBN 9780199233991

[12] Lyons, R.G. (2001). "Understanding Digital Signal Processing". *Prentice Hall*, ISBN 0-201-63467-8, p. 304.

[13] Juan J., Burred J.J. (2003). "An Objective Approach to Content-Based Audio Signal Classification". *Diplomarbeit eingereicht*.

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

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.