Digital system for sound sources location problem in multidimensional space using machine learning methods

Digital system for sound sources location problem in multidimensional space using machine learning methods

Maciej Walczyński[a]*, Piotr Ruszkowski, Wojciech Bożejko[a]

[a]Faculty of Electronics, Wrocław University of Science and Technology,

maciej.walczynski@pwr.edu.pl

Abstract. The issues presented in this work relate to the possibility of detecting the location of the sound source and its effectiveness in production conditions using recordings made with artificial head and processed with using an artificial neural networks. The use of an artificial head and artificial neural networks was motivated by an attempt to map human perception using available computer technology. The work attempts to map the features used for location human sources of sound through digital signal processing and machine learning. Artificial neural networks are commonly used, among others, in image recognition, where, using a camera or camera, the algorithm classifies objects in the image. Machine learning algorithms in addition allow for the implementation of self-learning speech or text recognition models. As part of the work, it was decided to localize the sound for a collection of fifteen different items sources relative to the head. To avoid duplicating similar measurements, the sound source was to the left of the artificial head. For signal processing and programming language was used to prepare the artificial neural network model Python with the Numpy numerical library, a library designed for signal processing Scipy, TensorFlow and Keras.

1. The first section in your paper

Industry 4.0 or the fourth industrial revolution is no longer just theoretical terms. They are a real element of the surrounding - and ever-changing - world. Industry 4.0 is a collective concept, provided that we can use practices such as the integration of intelligent machines and systems.

The research problem posed is whether it is possible to apply machine learning algorithms in the process of automation of sound source localization to the challenges posed by the next industrial revolution.

The presented method of detecting the direction of sound wave propagation using artificial neural networks can be used to detect damage to the production socket. In this case, in combination with the classification methods it is possible to indicate the type of failure and the machine to which it relates [1,2].

Audio signals collected in production halls - which are places with non-laboratory acoustic conditions - in addition to the component carrying useful information very often contain noise [3,4]. That noise component can be reduced using adaptive filtering algorithms [5, 6].

The proposed solution may find practical application in safety systems during the detection of a voice event (e.g. employee shouting) to designate a machine that should be automatically stopped in the event of a safety hazard or an accident at work [7].

2. Acoustic signals – theoretical background

2.1. Digital signal processing

Acoustic wave is a pressure disorder spreading in a spring center; the particles of the centre where the wave propagates are leaning out of the balance position, transmitting mechanical energy to other

particles. An acoustic wave can be received by a human ear and converted into an electric signal using a microphone [8, 9]. An analog sound signal, such as an acoustic wave-stimulated microphone, can be converted to digital form. The analog-to-digital conversion scheme is shown in Figure 1. The first step in conversion is low-pass filtration which leads to eliminate component frequencies above half of the sample rate of the transducer according to Kotielnikov-Shannon's assertion. If you do not use an anti-aliasing filter, you may be leaking the spectrum. The filtered signal must be subjected to discretion in the field of time; every specified quantum of time time τ ($\tau = 1/f_p$, where f_p - sampling speed), the peak value of the analog signal is measured. After the signal is discriuscated, the measured peak value is quantified in the time field. The voltage value of each sample is represented by a discrete set of values, the size of which depends on the bit resolution of the transducer. As the bit scan increases, the dynamic range that can be recorded with a given transducer increases [10, 11].

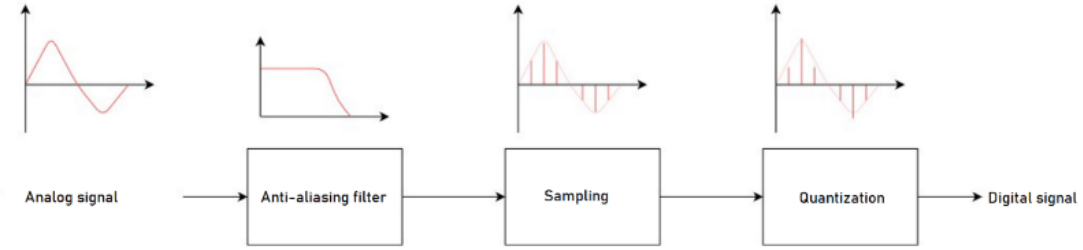


Fig. 1. Schematic diagram of analogue-digital processing.

2.2. Correlation function

Cross-correlation function (marked as R_{xy}) is a measure of similarity between two signals. The method of calculating the correlation function for discrete signals shows the Formula 2.1, where $x[n]$ and $y[n]$ are input signals, while N is the length of the x signal.

$$R_{xy}(k) = \sum_{n=0}^{N-1} x[n]y[n-k] \qquad (2.1)$$

2.3. Signal envelope

The calculation of the signal envelope can be achieved using hilbert transfromation marked with the symbol $H(x)$. To receive an $x[n]$ signal boundary, calculate the module of this transformation. The calculation of hilbert's discrete transformation shows a pattern [12]:

$$H(x[n]) = h(k) = \begin{cases} \frac{2}{\pi}\sum_n \frac{x(n)}{k-n}, k \text{ and } n \text{ odd} \\ \frac{2}{\pi}\sum_n \frac{x(n)}{k-n}, k \text{ and } n \text{ even} \end{cases}. \qquad (2.2)$$

2.4. Fourier transformation

Fourier transformation $F(x)$ is an important mathematical tool that allows you to calculate the spectrum of the test signal. The digital signal processing uses the *Discrete Fourier Transform* (*DFT*) or its optimized implementation called *Fast Fourier Transformation* (*FFT*). To calculate the continuous signal spectrum, use the classic Fourier transformation presented as follow:

$$F(x) = X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt. \qquad (2.3)$$

The equation form of formula 2.3 for discrete values (digital signal) is as follows:

$$F(x) = X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi}{N}kn}, \qquad (2.4)$$

for $k = 0,1,2, \dots, N-1$.
The formula given in 2.4 equation is called *Discrete Fourier Transform.*

3. Perception of the direction of sound by human

Locating the sound source by humans and animals is a very important skill; it allows us to indicate the direction of the object or danger we are interested in. Based on information about the location of the acoustic wave source, we can focus our eyes quickly on it [13]. The location of the sound, i.e. the assessment of the distance of the acoustic wave source and the direction from which this wave occurs on the basis of so-called location factors. The location is based on one-side dissionable signal processing and on the basis of a comparison of signals reaching both ears. Location factors depend on the physical properties of the sound. We can analyse factors depending on whether the signal is sinusoidal agitation with constant frequency and amplitude (simple tone), or composite signal with variable component frequencies and modulated amplitude. In addition, the usefulness of specific location factors also depends on the direction of the sound investigation. In the horizontal plane shown detection is very precise (reaching 1˚ accuracy). The resolution of the location in the middle or front plane is smaller (approximately 4˚). Horizontal detection is carried out primarily using the access time difference (*ITD*) and on the basis of the inter-ear level difference (*ILD*). In the case of front and middle detection, the relevant information is the signal spectrum shaped by human auricles and reflections from the torso [13]. For sinusoidal stimulation with constant amplitude and frequency, the source of which is located at a non-zero angle θ sound will first reach the ear closer to the source at the time of t_0 with a sound pressure level p_0, then reach the other ear after $t_0+\Delta t$ time and with a sound pressure level $p_0-\Delta p$. Δt and Δp values are binaural location factors. The difference in the time of the handle is determined by the inter-ear time difference (*ITD*), the difference in the level of acoustic pressure is referred to as the inter-ear level difference (*ILD*).

4. Artificial neural network model

The main part of the system for sound sources location problem in multidimensional space is the input signal feature extraction module and the artificial neural network model. It is not possible to design and train an artificial neural network model without in-depth analysis of the issue. In order to understand the important dependencies, numerous recordings and measurements were made to allow the problem to delve into, as well as to subsequent evaluation of the algorithm. The main issue related to the assumptions of the work was the mapping of psychoacoustic phenomena, used to locate the sound source using a computer algorithm. The measurements described in the subsection have made it possible to determine the relevant characteristics on the basis of which it is possible to detect the direction of acoustic wave investigation into the artificial head.

5. Experiments

The main issue related to the assumptions of the work was mapping phenomena psychoacoustic, used to localize the sound source using a computer algorithm. Performed measurements described in the subsection allowed to determine the essential features on the basis of which it is possible to detect the direction of the sound wave approach to artificial head.

5.1. Preparation of test signals

An important part of the study was the preparation of a sequence that was reproduced by the source. For the possibility of analyzing the transmittance of the entire electroacoustic system, a sine wave signal with logarithmically over tuned frequency was placed at the beginning of the sequence. In order to train and verify effectiveness, speech signals and broadband signal containing a large amount of information were placed in the sequence. Figure 2 shows the spectrum of selected broadband signals used in the database, as we can see this signal is a broadband signal with significant unevenness. In Figure 3 and 4, we can observe the spectre of selected speech signals which were elements of the test string Figure 3 represents the spectrum of male speech, while Figure 4 shows the spectrum of the female speech signal. The broadband signal is a signal with a wider spectrum than a speech signal.

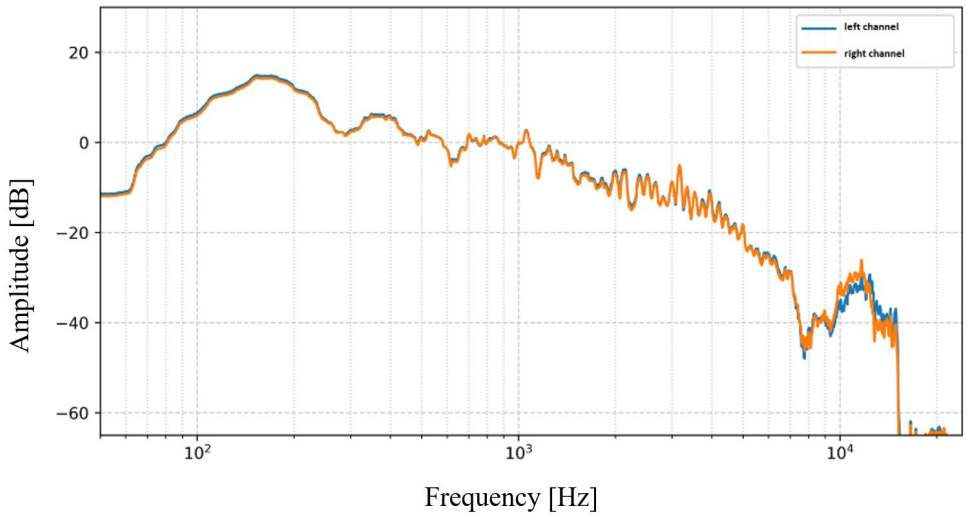


Fig. 2. Spectrum plot of an example of broadband signal containing a large amount of information. The component amplitude was referenced to a level for the 1 kHz frequency.

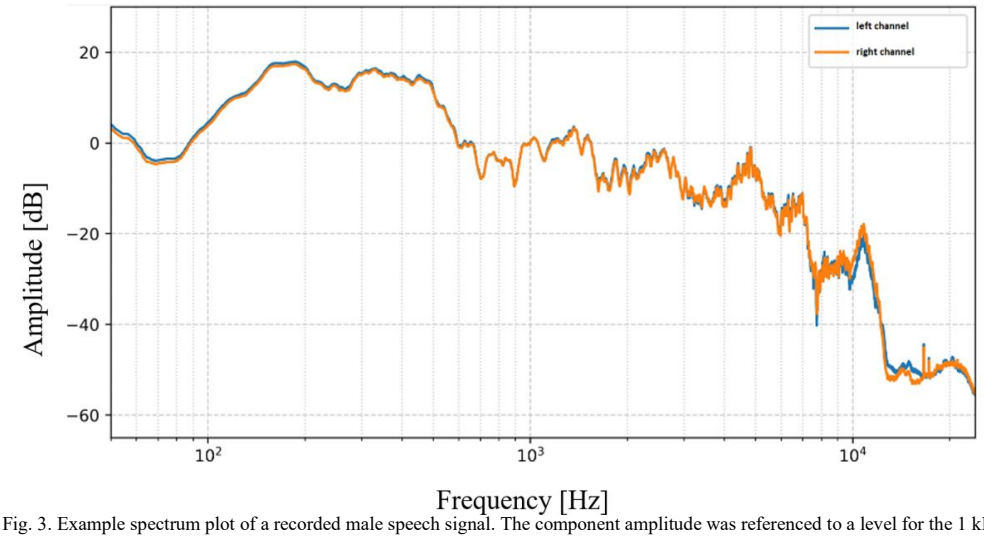


Fig. 3. Example spectrum plot of a recorded male speech signal. The component amplitude was referenced to a level for the 1 kHz frequency.

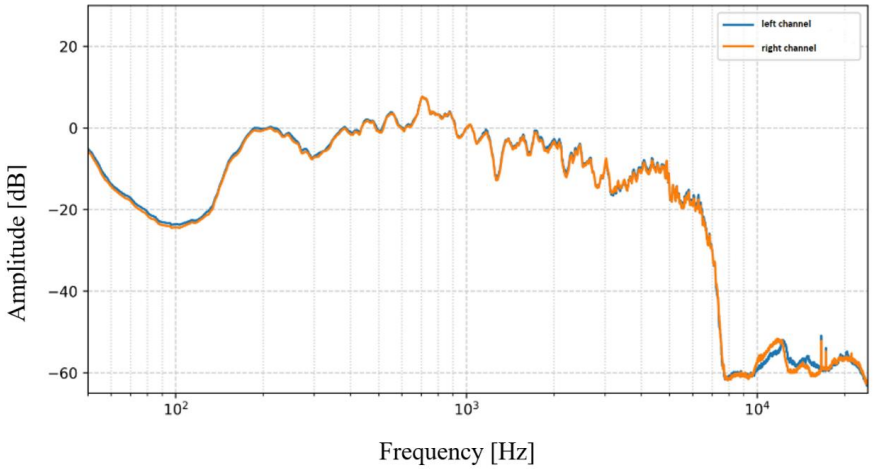


Fig. 4. Example spectrum plot of a recorded female speech signal. The component amplitude was referenced to a level for the 1 kHz frequency.

In the case of speech signals, we can see that the spectrum of male speech is shifted towards small frequencies in relation to the female speech spectrum. In the presented spectrum, we can observe that each signal contains relevant information (drop not exceeding 20 dB) in the range of about 200 Hz, up to a frequency of about 3 kHz. The prepared string consisted of twenty-eight samples exported to a single wav file, with a sampling rate of 48 kHz and a resolution of 24 bits. When preparing a test string, a similar speech signal content was sought to the broadband signal. The samples were separated by a silence lasting two seconds.

Table 1. Content of the test signal

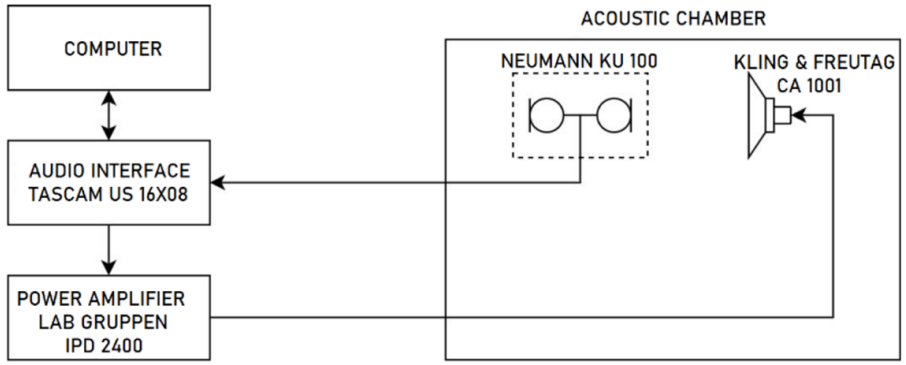
Type of the signal	Number of items
Test signal (sweep)	1
Broadband signal containing a large amount of information	12
Female speech signal	7
Male speech signal	8

5.2. Recordings of the sample database

The object in which it was decided to make the recordings was an acoustic chamber designed to simulate the free field, located in building C-16 at the Wrocław University of Science and Technology. The chamber has a reflective concrete floor and a strongly absorbing wall and ceiling. The reflective surface of the concrete floor allowed additional information to be recorded in the recorded signal resulting from comb filtration resulting from direct signal interference and reflected signal. Due to the high acoustic absorbency of the walls and ceiling, it was considered that they did not introduce additional apparent sources.

According to the assumptions set out in previous chapter, a speaker device was used as an acoustic wave source. Figure 5 shows a diagram of the connections of the electroacoustic track used to prepare test signals.

The source of the audio signal was a computer with Reaper software; the digital phononic signal was converted to an analog signal using the TASCAM US 16x08 interface. The interface first output was connected to the input of the first power tip channel, the output of which is connected to the input of the speaker device.
The acoustic signal emitted by the speaker device was recorded using an artificial Neumann head connected to the first two microphone inputs of the TASCAM external sound card and recorded in



reaper. A s, which allows t

Fig. 5. Wiring diagram when recording test samples

The measurement consisted in recreating the prepared test signal and simultaneous recording of the signal from both microphones constituting the construction of an artificial head. A total of 15 different speaker device arrangements were made relative to the artificial head.
Figure 6, shows the reference system adopted; the *R* value is the distance between the artificial head and the plane on which the sound source was located. *H* describes the height of thespeaker device above the ground, measured relative to the center of the speaker device. The angle α describes the angle between the head and the speaker device on the horizontal (horizontal) plane.

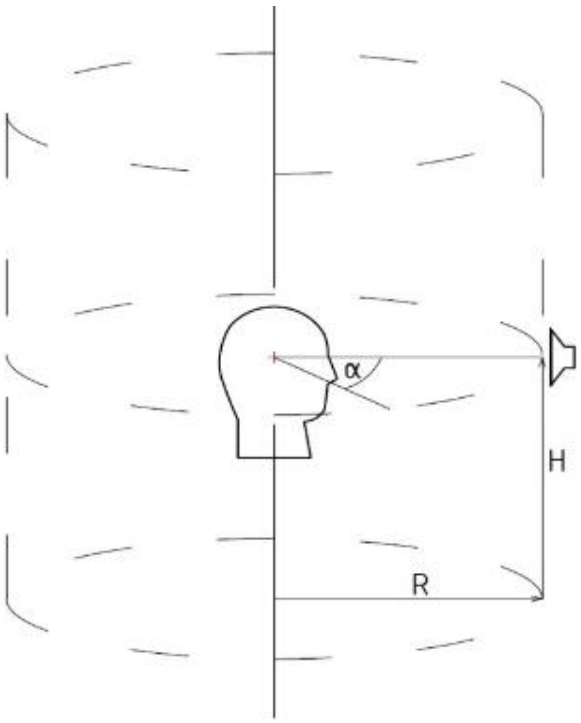


Fig. 6. The reference system adopted by the authors in cylindrical coordinates.

The artificial head was placed so that the ear canal is at an altitude of 1.70 m relative to the ground. During the measurements, the height of the *H* speaker device depended on the position, for the lowest position was 1.40 m for the highest 2.0m. For horizontal relative to head measurements, the sound source was 1,70 m above the ground so that it was at the height of the artificial head. The small value of changes

in the facade of the speaker device is due to technical limitations; speaker device has always been perpendicular to the ground (it is not possible to change the alignment of the radiation plane), so it was necessary to set the device so that the artificial head does not fall outside the radiation area of the device. As we can see in Figure 6, in order to ensure the most even frequency characteristics possible, the facade angle could not exceed about 15˚.

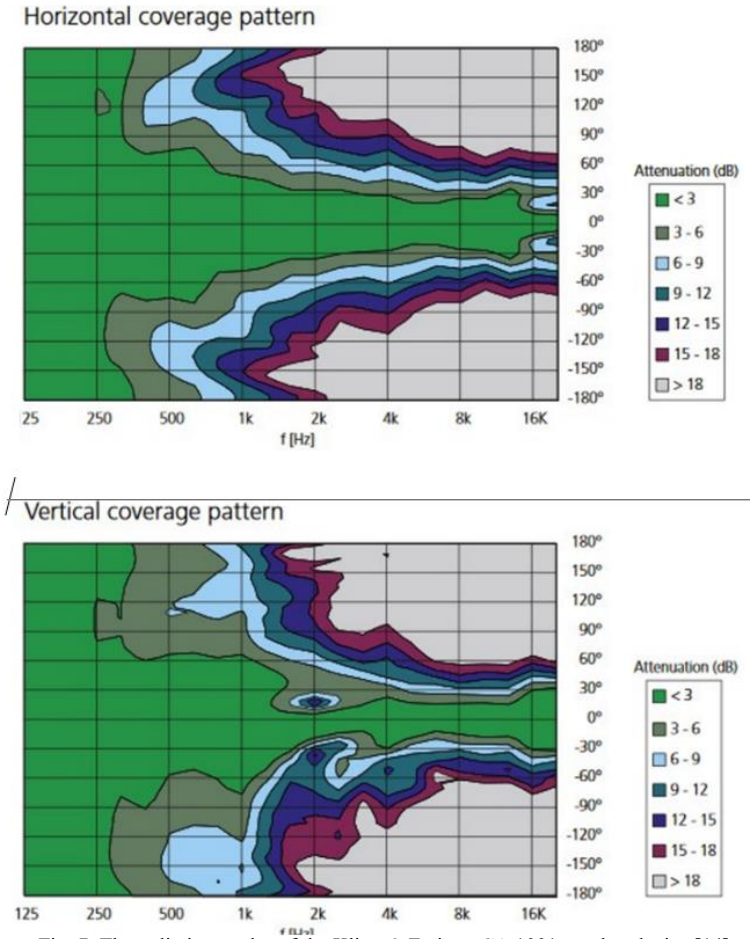


Fig. 7. The radiation angles of the Kling & Freitag CA 1001 speaker device [14].

Since the recording in each position was a single wavfile, it was necessary to divide them into smaller, shorter signals containing individual fragments of the entire signal. To facilitate work related to the organization of signals on the disk and their subsequent processing, the position and signal number are encoded in the file name. In order to be able to easily expand the recording base in the future, it was decided to encode also the distance between the signal source and the artificial head. The following file naming convention was adopted: dXXX-YYY-ZZZ-N.wav, where 'XXX' was the *R* distance, given in centimeters (currently all files begin with the 'd150' string), 'YYY' is responsible for information about the height of the *H* device, while 'ZZZ' is responsible for the angle on the horizontal plane expressed in degrees (000, 045, 090, 135, 180), at the end ('N') there is a sample number marked with a natural number, the signals have been numbered from 0 to 27.

The entire track transmitting for the first position adopted is shown in Figure 8. The sample values were standardized against the value for $f = 1kHz$, in addition, the graph was smoothed, for this purpose a rectangular window with a length of 200 samples. As expected, the graph for the left and right canals

overlap to a very large extent, both ear of an artificial head, are distant from the same distance. The track transmittance was largely influenced by a speaker device whose frequency response is Figure 8.

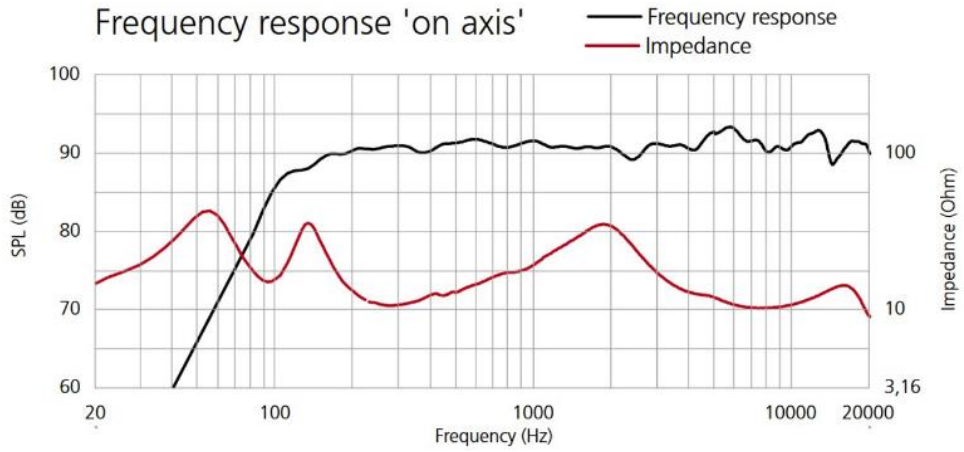


Fig. 8. Frequency response of the Kling & Freitag CA 1001 speaker device [14].

5.3. Analysis of recorded material

Monaural localization factors

On the basis of the assumptions presented, an algorithm was to be prepared to extract selected characteristics from recorded sound samples. The artificial neural network input vector can be represented as follows:

$$X = [ILD, ITD, T_{1L}, T_{1R}, T_{2L}, T_{2R}], \qquad (5.1)$$

where *ILD* and *ITD* are values determined by the analysis of the differences between the signal reaching the left ear and the signal reaching the right ear. T_{1L}, T_{1R}, T_{2L}, T_{2R} values refer to signal levels in selected tertiary bands. The exact way in which the features are extracted is described below.

The value modeled after ILD was calculated as the difference in the effective signal value recorded by the left ear and the effective signal value recorded by the right ear. Since the expected effective values of both signals are very small, it was decided not to express values in decibels, a simplified method of calculation can be expressed with the formula:

$$ILD = \sqrt{\frac{\sum_{n=0}^{N-1} x[n]^2}{N}} - \sqrt{\frac{\sum_{n=0}^{N-1} y[n]^2}{N}}, \qquad (5.2)$$

where $x[n]$ is the vector of left channel samples, $y[n]$ is the signal sample vector recorded on the right channel, N is the signal length in the samples.

The implementation normalizes the effective value of both signals relative to the signal received by the left ear; this treatment was carried out due to the subsequent processing of values by the activation block in artificial neuron.

The calculation of the inter-ear time difference of access was made based on the cross-correlation function of the modules of the Hilbert transform of both signals. For calculated $R_{gh}(k)$, where $g[n]=|H(x[n])|$ and $h[n]=H(y[n])|$, the value k for which the correlation function reaches its maximum is found, this value is the delay value in the samples between the signals. Because an artificial neural network accepts a delay expressed in milliseconds as an input value, the delay in the samples is multiplied with a sampling period of $^1/_{48} \cdot 10^{-3}$ s.

Binaural localization factors

The extraction of single-ear location factors was an important element to increase detection efficiency after the introduction of samples recorded for $H \neq 1.7m$. The mono-usable factors are due to the function of head transmitting, and therefore a characteristic analysis has been carried out in the field of frequencies.

Figure 9 shows the transmittance of the entire electroacoustic system when the source is positioned in front of the artificial head ($\alpha = 0^{\circ}$ and $H = 1.7m$). As we can observe, the spectrum has local minima and unevenness that result from the transmission of the head, input signal, transmission of the speaker device and phenomena such as comb filtration resulting from the interference of the wave reflected from the direct wave substrate; in view of the complexity of the electroacoustic track, finding important and interesting properties resulting directly from the function of head transmitting, is a difficult task.

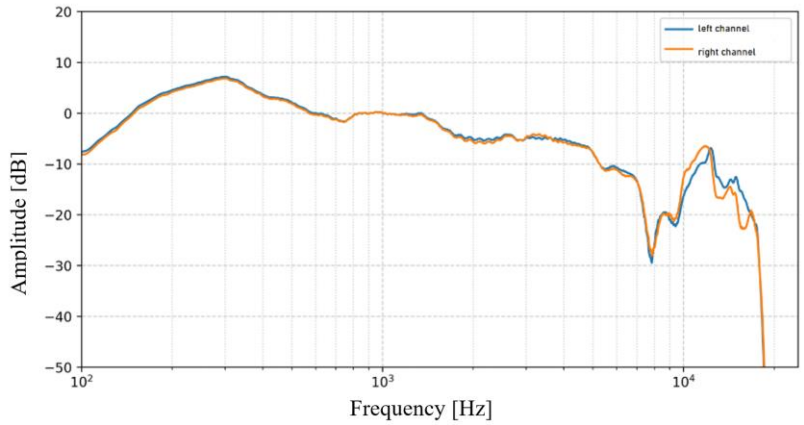


Fig. 9. System transmittance, H = 1.7 m, α = 0 °.

It was decided to analyze the signal level in two different tertiary bands. The tertiary bands were selected in an experiential manner. Octave bands may have included such significant changes in their width that the level differences in the sound source location function could be too small to provide important information for the developed model of the artificial neural network. Critical bands resulting from the distribution of deviations on the underlying membrane of the cochlea are narrower than the tertiary bands, so there was a certain likelihood that signals with a spectrum similar to the speech signal might not contain information at a significant level in selected bands. When choosing a series of bands in which the average speech and musical signal contains component frequencies [13, 15, 8] it was important to choose bands in which the average speech and broadband signal containing a large amount of information contains component frequencies [13, 15, 8].

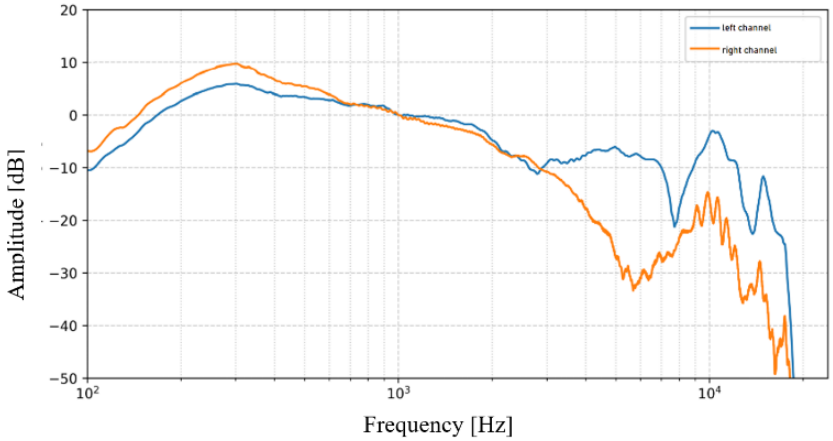


Fig. 10. System transmittance, H = 1.4 m, α = 90 °.

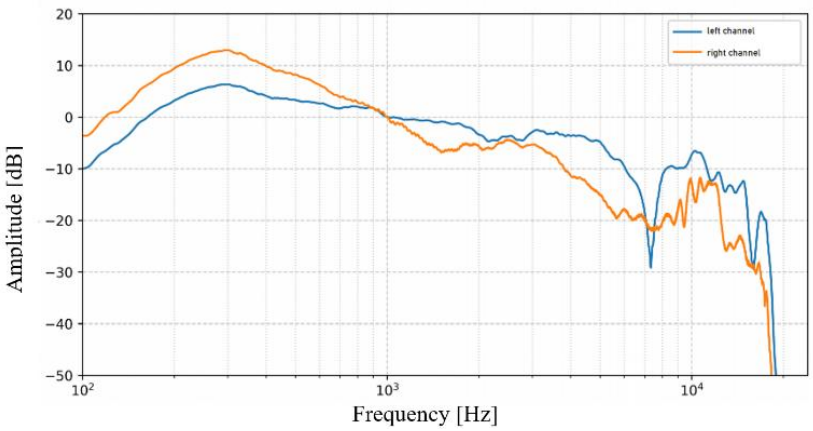


Fig. 11. System transmittance, H = 1.4 m, α = 45 ˚.

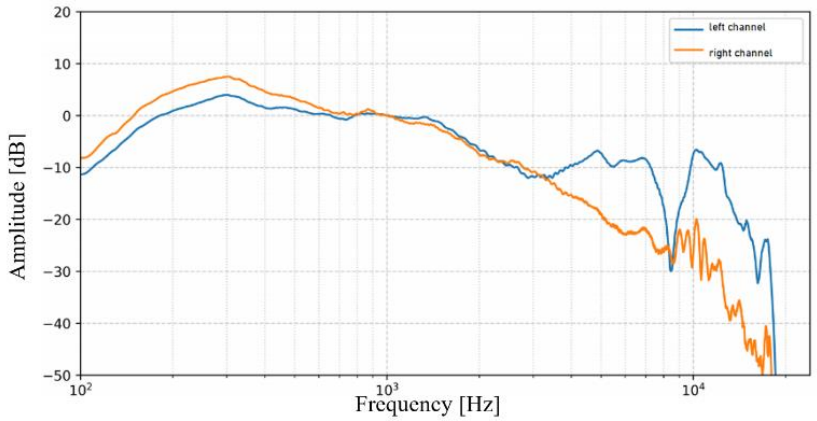


Fig. 12. System transmittance, H = 1.7 m, α = 90 ˚.

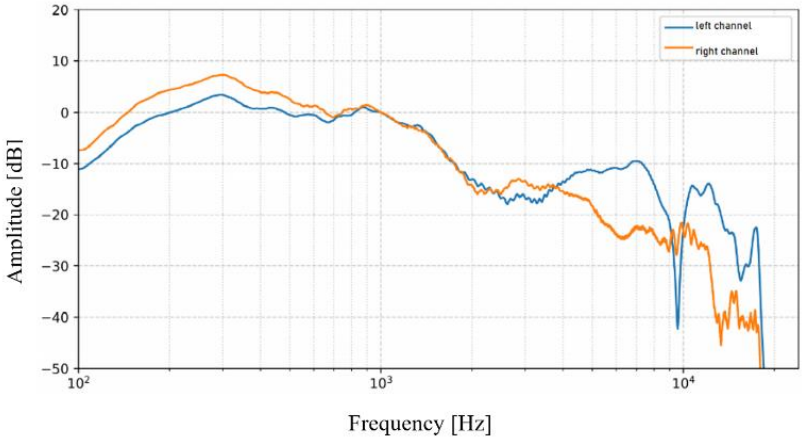


Fig. 13. System transmittance, H = 2.0 m, α = 90 ˚.

As we can see, in Figures 10 - 13, there are large differences in the sound level between the left channel and the right channel, while the ratio of these levels depends on the facade and the angle on the horizontal plane α. Based on spectral analysis and the common band part described in previous section, the extraction of features is based on the sound level of 400 *Hz* and the 2 *kHz*. In order for the characteristic values per signal to be constant in the effective signal level function at the source position, the sound level in each period is given relatively relative to the pressure level value for the 1kHz frequency.

5.4. Implementation of an artificial neural network

Based on the *Keras* library using the *Tensorflow* package, the artificial neural network model was prepared. The size of the input layer is determined by the prepared number of features, while the size of the output layer is determined by the given classes; i.e. the source positions relative to the artificial head [15, 16,17,18].

The network had six input neurons and fifteen output neurons. The function of activating the output layer was the softmax function [19]. The number of hidden layers and their size have been selected empirically based on numerous iterations to improve the effectiveness of the model. The final network model has two layers hidden, each with 64 neurons, as activation functions, a hyperbolic tangent was used.

Between each hidden layer, there was a dropout layer. The first dropout layer had a coefficient of 0.1, while the second rate was 0.3.

Table 2 Description of the structure of the designed network.

Layer type	Number of neurons	Activation function
Input layer	6	---
1-st hidden layer	64	Tanh
2-nd hidden layer	64	Tank
Output layer	15	Softmax

A thousand generations were made, three hundred iterations consisting of strings containing fifteen samples were made in each generation; the same signal in all fifteen positions. Three hundred samples were used for training, consisting of twenty different signals, each signal was emitted in fifteen positions.

The trained model has been saved to be able to move it to the finished application and further study its effectiveness. The prepared model was verified for effectiveness on a group of hundred and twenty test samples that did not contain within the teaching string. The test string contained eight different signals, each recorded in fifteen positions. Due to the randomness of sampling selection, the exact content of the signals from both strings could not be determined. The effectiveness of the model has 88.3%.

In order to speed up the process of network learning and object classification, feature extraction was held once for all signals. An algorithm has been prepared, which on based on the path given by the user, searches for all signals in wav format, and their analysis. The results of the calculations are transferred to a text file, so that they can be quickly recalled for training purposes later on networks.

6. Summary

Mapping the methods used by humans to detect the location of a sound source in space using digital signal processing techniques and artificial neural networks allowed for direction detection with very high efficiency reaching almost 90%, in an environment in which a learning string was prepared.

Using python programming language and relevant libraries, it allowed to achieve the necessary calculations in a relatively simple way. The most complex computational module is the module responsible for calculating the delay between two signals.

The shape of the auricles and head significantly affects the transmitting of the track, so it is possible to distinguish the direction from which the acoustic wave occurs.

Detection of the location of the sound source by the algorithm using an artificial head, is feasible for selected assumptions.

Despite the impact of transmitting the entire combined electroacoustic track preceding the artificial head, it was possible to analyze the essential characteristics of the signal and evaluate the artificial neural network.

The use of machine learning allows us to fine-tune the designed application to work in specific acoustic environments.

Despite the small facade angle resulting from hardware limitations, the model was able to achieve satisfactory performance.

In the future, a number of experiments are planned to be carried out under non-absorbent conditions. In production halls, where the noise level is much higher, the system should be equipped with adaptive algorithms to reduce this unwanted noise [4]. In addition, for greater efficiency, the implemented algorithms can be aligned and run on a GPGPU computing cluster, for example [10,20,21].

References

[1] B. Rubhini, P. Vanaja Ranjan, "*Machine condition monitoring using audio signature analysis*", Signal Processing Communication and Networking (ICSCN) 2017 Fourth International Conference on, pp. 1-6, 2017.

[2] T. de M. Prego, A. A. de Lima, S. L. Netto, E. A. B. da Silva, "*Audio anomaly detection on rotating machinery using image signal processing*", Circuits & Systems (LASCAS) 2016 IEEE 7th Latin American Symposium on, pp. 207-210, 2016

[3] A. Dobrucki, P. Pruchnicki, P. Plaskota, P. Staroniewicz, S. Brachmański and M. Walczyński (July 20th 2016). "*Silent Speech Recognition by Surface Electromyography*", New Trends and Developments in Metrology, Luigi Cocco, IntechOpen, DOI: 10.5772/60467.

[4] M. Walczyński, D. Ryba (2019). "Effectiveness of the acoustic fingerprint in various acoustical environments". 137-141. 10.23919/SPA.2019.8936781.

[5] A. Dobrucki, W. Bożejko, M. Walczyński, "*Parallel LMS-based adaptive algorithms of echo cancellation*", Computer Science2014 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) 2014

[6] A. Dobrucki, M. Walczyński, W. Bożejko, "*Family of parallel LMS-based adaptive algorithms of echo cancellation*". Computational Methods in Science and Technology. 2015, vol. 21, nr 4, pp. 191-200

[7] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, F. Piazza, (2015). "*An integrated system for voice command recognition and emergency detection based on audio signals*". Expert Systems with Applications. 42. 10.1016/j.eswa.2015.02.036.

[8] I. Corbett, "*Mic It!*", Focal Press, 2015. ISBN 9780415823777.

[9] A. Dobrucki, "*Przetworniki Elektroakustyczne*" (in polish). Wydawnictwa Naukowo-Techniczne, 2007. ISBN 9788320432145.

[10] A. B. Downey, "Think DSP: Digital Signal Processing in Python (1st. ed.) ". O'Reilly Media, Inc

(2016).
[11] R. G. Lyons, "*Understanding Digital Signal Processing (1st. ed.)*". Addison-Wesley Longman Publishing Co., Inc., USA. 1996
[12] S. Kak, "*The discrete hilbert transform*". Proceedings of the IEEE, vol. 58, pp. 585-586, April 1970.
[13] B. Moore, "*Wprowadzenie do psychologii słyszenia*" (in polish). Wydawnictwo Naukowe PWN, 1999. ISBN 9788301127664.
[14] Kling & Freitag. Technical manual k&f ca 1001, URL https://www.kling-freitag.com/content/uploads/ds_ca-1001_en.pdf
[15] R. Venkatesan, B. Ganesh, "Full sound source localization of binaural signals". 2017.
[16] Keras library documentation. "*Keras: The python deep learning library*", July 2019. URL https://keras.io/.
[17] V. Zhou, "*Keras for beginners: Building your first neural network*", URL https://victorzhou.com/blog/keras-neural-network-tutorial/.
[18] N. Ketkar, "Introduction to Keras. In: Deep Learning with Python". Apress, Berkeley, CA
[19] Gao, Bolin & Pavel, Lacra. (2017). "On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning".
[20] Bożejko W., Dobrucki A., Walczyński M., "*LMS algorithms parallelization in GPGPU environment"*, Elektronika, 2011, R. 52, nr 5, 49-53.
[21] Walczyński M., Bożejko W., "*Noise reduction with using parallel algorithms*", Noise Control '10, ISBN 978-83-7373-077-9.