Acoustic signal processing via neural network towards motion capture systems

E. Volná, M. Kotyrba, R. Jarušek
Department of informatics and computers, University of Ostrava, Ostrava, Czech Republic

Abstract - *The aim of this article is to outline possibilities of sound and its physical properties during shooting of moving objects. Attention was devoted to the specific location of a fixed point in the space and time. We present two proposed methods that are based on neural networks. We also proposed appropriate topologies of the systems that depend on the required accuracy, acoustic properties and selected sound technologies. At first, we identified a distance between an active transmitter and a receiver on the basis of sound pulses transmitted from transmitters in the defined domain. After that a neural network uses obtained distances between transmitters and a receiver as its inputs to determine an actual position of the receiver in space. We developed two models, which outcomes are compared in conclusion.*

Keywords: Acoustic signal processing, neural networks, motion capture system, Fourier transform.

1 Sound waves processing

When sound impacts on the solid barrier, it causes its reflection or bending which depend on the ratio between the size of the barrier and the wavelength of sound. If the dimension of the barrier is bigger than the wave length of the sound, the sound is reflected according to the rule: *"The angle of reflection equals the angle of incidence"* and this phenomenon can be simply viewed as the problem of propagation of light rays. Value of intensity (residual energy) of reflected sound signal is defined by the physical properties of the material and it is different for different sound frequencies. Generally speaking, for the lower frequency absorption coefficient is smaller, with increasing frequency coefficient of absorption is increasing. We write (1):

$$a = \frac{i_0 - i}{i_0} \qquad (1)$$

where:
a - sound absorption coefficient at reflection
i - intensity of the reflected waves
i_0 - intensity of the incident wave

Fig. 1 shows the sound pulse as a rectangular signal, which is generated from the sum of odd harmonics frequencies with a prescribed amplitude. Additionally, it is very easy generated and its transmission over sinusoidal signal is multiple. Just these sound waves form the basis of motion capture systems that are aims of this article.

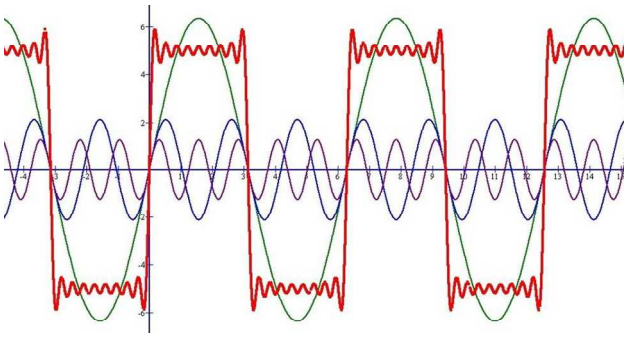

Figure 1: A sound pulse as a rectangular signal [11]

2 Acoustic motion capture systems

Capturing motion or motion tracking (MoCap) is used to provide a digital recording using the markers. Currently, there are several techniques for tracking. Computer software which is provided to the motion capturing record positions, angles, velocity, acceleration, and pulse points in the real time. For now, an unused option of the Motion Capture is a system for determining the positions of points in the space which uses the physical properties of audible sound. Since the speed of sound propagation in the environment is constant, it's possible to calculate an audio signal's absolute distance according to the degree of its delay. If this happens for at least three transmitters, receivers can determine the position of the spatial coordinates via triangulation. Several motion capture technologies have been proposed in the last two decades. The advantages and disadvantages of the dominant approaches are argued in several excellent surveys [3, 5].

Acoustic systems use the time-of-flight of an audio signal to compute the marker locations. Most current systems are not portable and handle only a small number of markers. With the Bat system [13], an ultrasonic pulse emitter is worn by a user, while multiple receivers are placed at fixed locations in the environment. A system by Hazas and Ward [4] extends ultrasonic capabilities by using broadband signals; Vallidis [9] alleviates occlusion problems with a spread-spectrum approach; Olson and colleagues [7] are able to track

receivers without known emitter locations. The Cricket location system [8] fills the environment with a number of ultrasonic beacons that send pulses along with RF signals at random times in order to minimize possible signal interference. This allows multiple receivers to be localized independently. A similar system is presented by Randell and Muller [10], in which the beacons emit pulses in succession using a central controller. Lastly, the WearTrack system [3], developed for augmented reality applications, uses one ultrasonic beacon placed on the user's finger and three fixed detectors placed on the head-mounted display. This system can track the location of the finger with respect to the display, based on time-of-flight measurements.

3 Acoustic motion capture systems based on neural networks

We present two proposed MoCaps that are based on neural networks, e.g. their appropriate topologies that depend on the required accuracy, acoustic properties and selected sound technologies. At first, we identified a distance between an active transmitter and a receiver on the basis of sound pulses transmitted in the defined domain. After that a neural network uses obtained distances between transmitters and a receiver as its inputs to determine an actual position of the receiver in space.

3.1 System design

The article introduces experimental study of an audible MoCap system developed via neural networks. Designing a measurement system has been defined the following initial requirements [12]:

- Active area (domain), where the captured objects move, has to be so large to be able to cover the range of moving objects.
- Active area should not restrict the moving objects.
- The system accuracy must be constant throughout the active area.
- The system must be able to adapt to environmental changes (e.g. change in temperature).
- The system must be able to detect measurement errors and correct them.
- The output of the system must be data that should be acceptable in other systems (e.g. 3D programs).
- The system should be able to work in real time.
- The whole system, including technology, should be applicable in any environment.

According to the initial requirements, we proposed two system topologies containing five or three transmitters positioned around the space. All transmitters were put into a horizontal plane so that the plane split the space into two half-space, namely the half-space above the floor and half-space under the floor. We introduced a coordinate system into the half space above the floor, see Fig. 2,3. Our proposed system is based on speakers that generate a signal that is recorded sensor. Gradually we emit an acoustic pulse from different transmitters into the microphone. As the space is defined with microphone placement transmitters, we are sure that one sound pulse leaves the room with a microphone even before then second transmitter in turn sends its pulse. Thus, in the area one pulse is only in the current time.

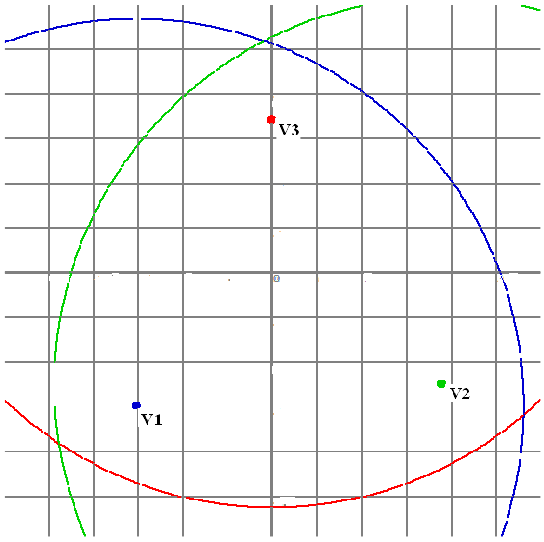

Figure 2: A coordinate system 3 transmitters' positions (V1 - V3).

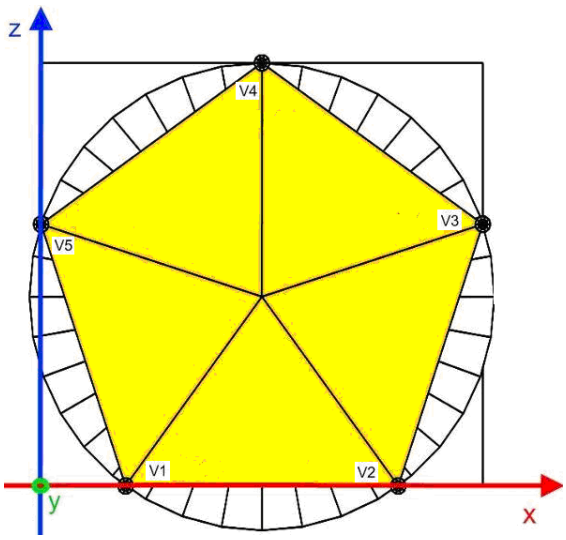

Figure 3: A coordinate system 5 transmitters' positions (V1 - V5).

We had to fulfill the following conditions of sound parameters in order to system worked well [6]:

- The system used sound waves at a frequency of 4410 Hz.
- Sound pulse, used as a measurement medium and it is radiated by any transmitter, must leave the domain before any other transmitter starts sending its impulses. This is the most important condition for the proper system functioning.

- Sound pulse must be adequately long to receive it the satisfaction in receiver and process it.
- Sound pulse must be adequately short not to overload space domain by reflections from walls or objects in the room.

There were made measurements in domains shown in Fig. 2,3 where we changed the receiver position for each measurement and we obtained 33 audio records. Initially, receiver was placed in the static points in space in order to cover the edge of the domain too. Then the receiver was moving so we recorded its dynamic movement in time. Recorded material was transferred to the stereo base (where the left channel contained impulses of transmitted V1 - V5 and right channel contained a record from the receiver) in order to create training and test sets of neural networks.

3.2 Sound wave identification

Each speaker sends a signal (Fig. 5), which is shifted by optional time interval to the remaining generators. By optimization we can achieve such detection which is not dependent on the size of the scanning space, because these signals are clearly distinguishable.

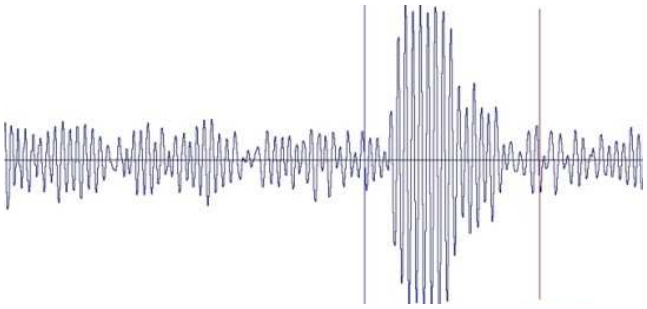

Figure 4: Non-filtered audible signal with environment noise

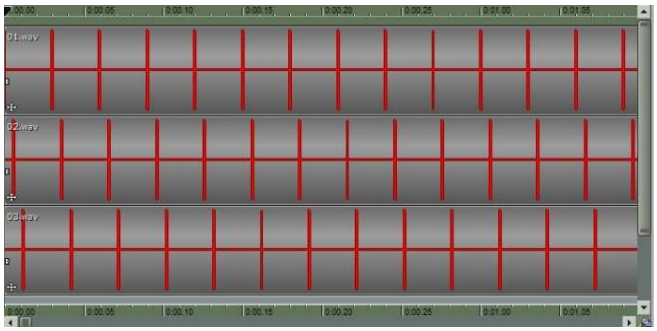

Figure 5: Shift of the individual audio signals from each speakers

To the filtered sound sample scanned by the receiver and to the detection the sound pulse's onset we use the Fourier transformation, specifically FFT - Fast Fourier Transform [1]. At 4410 Hz sample rate (set to sound card) and the number of samples 1024 (2^{10} - necessary for FFT) scanned sample is then processed by the transformation matrix and there is selected only zone with frequency of the sound pulse (4410Hz). A band remains unchanged, while the other zones are reset. Then we perform the inverse FFT and after that we get a filtered sample (Fig. 4). In such a filtered sample we simply find the maximum, which then determines the onset of the sound pulse in the sample.

This neural network is able to find the beginning of the sound pulse of transmitter and transform this information into a numerical value expressing the distance between the transmitter and receiver. We used a multilayer neural network with one hidden layer that was adapted by backpropagation algorithm [2]. Input data of the training set included fixed range of values of one sample with the length of one (main) sequence, which contained 882 patterns. Number of patterns in the training set was 1744. Neural network architecture is the following: 88 units in input layer, 120 units in hidden layer, 44 units in output layer.

Input vector of the training set included 88 values from the interval <0, 1>. Values present standard maximal and minimal subsequence values of 20 samples from the main sequence, e.g. pairs of maximum from positive numbers and minimum from negative numbers. The last two samples from the main sequence were omitted. Output vector of the training set included 44 values from the set {0, 1}. If we divide the main sequence into 44 parts (each part includes 20 samples), then the part, which contains a front edge flag of the pulse equals 1 and all other values remain this value.

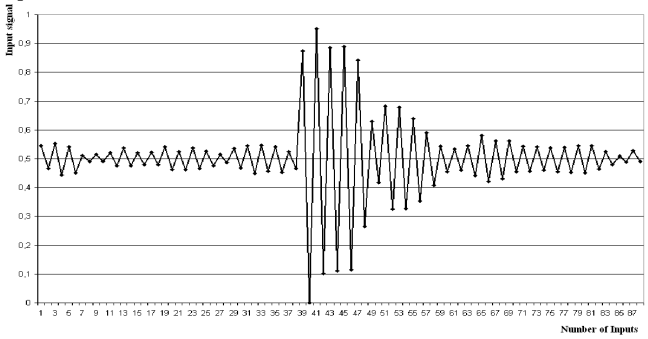

Figure 6: Non-filtered audible signal with environment noise

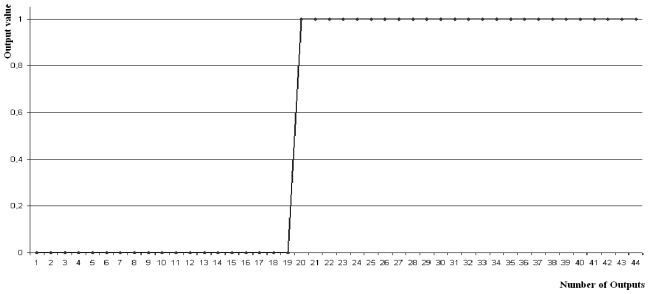

Figure 7: Visualization of the output training vector

Choice of format of input data (input vector) was an important moment, see Fig. 6. We preferred maximal and minimal values of subsequences, because their average values did not give desired results. Similarly, the format of output data (output vector) was proposed as a no decreasing function

with the skip point in front edge flag of the pulse (Fig. 7). Fig. 8 shows calculation of random sequences that form the test set.

The proposed network was able to recognize from input data the pulse signal with an accuracy of 20 samples (e.g. $20 * 0,\overline{7}\ \text{cm} \doteq 15,4\ \text{cm}$), which is higher than the level of our desired accuracy.

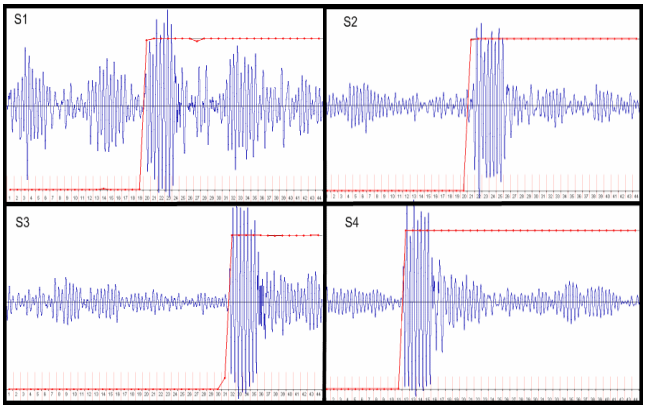

Figure 8: Test sequences (S1 - S4)

3.3 Coordinates generation

In our experimental study, we used a multilayer neural network with one hidden layer that was adapted by backpropagation algorithm [2] for the task of calculating the coordinates of points in space.

The philosophy of the application is simple. The distance between the individual transmitter and receiver is calculated from given coordinates of three or five transmitters and randomly generated three-dimensional coordinates of fictional receiver, which is located in the domain Fig. 2, 3. We must transform these values to the coordinates (*x*, *y*, *z*). Both data represent a training set which are used during a neural network adaptation. Each training pattern consists of three or five input components (the distance from three transmitters to a receiver) and three output components (*x*, *y*, and *z* coordinates in space). The actual distance is then determined by Euclidean distance calculations.

Experimental setting – 3 transmitters' positions

The neural network was adapted by set of 3000 training vectors, whose uniformly cover the all domain space (Fig. 2, 3). The suggested parameters of our experimental work are the following:

- Input layer: 3 units
- Hidden layer: 6 units
- Output layer: 3 units
- Activate function: a sigmoid
- Learning rate: 0,3

Experimental setting – 5 transmitters' positions

The neural network was adapted by set of 4000 training vectors, whose uniformly cover the all domain space (Fig. 2). The suggested parameters of our experimental work are the following:

- Input layer: 5 units
- Hidden layer: 12 units
- Output layer: 3 units
- Activate function: a sigmoid
- Learning rate: 0,3

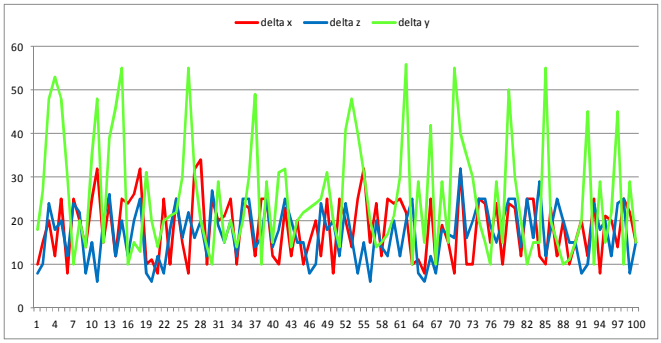

Figure 9: Measurement results - neural network error in cm (axes x: 100 test sequences). 3 transmitters' positions

3.4 Coordinates generation

In test phase, we used the adapted neural network for real data which were obtained from an audio sample. Of course it is necessary to normalize this data and because of it we determine the maximum distance at which the receiver (microphone) can occur. Distances are normalized to the interval <0, 1>.

Test set includes 100 patterns. Measurement results were shown in Fig. 9 (3 transmitters' positions) and Fig. 10 (5 transmitters' positions). Both experimental results are very similar. We are able to summarize them as follows:

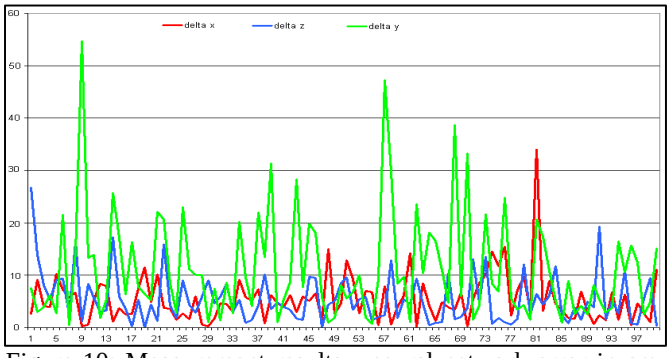

Figure 10: Measurement results - neural network error in cm (axes x: 100 test sequences). 5 transmitters' positions

- Calculating accuracy of *horizontal* coordinates (*x*, *z*) was, on average, 2,5 cm.

- Calculating accuracy of the *vertical* coordinate (*y*) was, on average, 5,5 cm. This reality was due to real disposition of transmitters, where the change about 1 cm in height indicated minimal changing of distance from transmitters. In the case that the vertical coordinate was close to zero, the network error was increased in the calculation.

4 Conclusion

The objective of the paper helps to outline the possibilities of using sound and its physical properties during shooting of moving objects in space and time for the purpose of converting these movements into virtual space. We found out that Motion Capture Systems using sound can be applied in real conditions, and physical properties of sound we can really use. Crucial component of the system are neural networks, thanks to their ability of generalization and information filtering, the system was allowed to process mixed and noisy data.

To solve data extraction from sound waves, we propose a new structures of training sets corresponding to the original structure that means it is used to separate all difficult recognizing patterns from the training data set, therefore the main emphasis of this paper is focused on the fact, how to properly design training set for given neural networks. This work deals with determining of receivers' positions in space and time. The proposed systems also solve specific moving objects. Here, the limiting factor is only a number of transmitters, the domain size and average acoustics properties in room. Number of receivers can be in this configuration theoretically unlimited, we have to provide sufficient computing power. We developed two models with 3 or 5 transmitters. Both models were compared and we received very similar experimental outcomes. As the vertical coordinate was close to zero, both models' errors were greater than in horizontal direction. For this reason, we are going to develop 3D MoCap system, which could be able to reduce inaccuracies in vertical direction too.

5 References

[1] Brigham, E. O. (2002). The Fast Fourier Transform. New York: Prentice-Hall.

[2] Fausett, L., (1994),: "Fundamentals of Neural Network". 1st ed. Prentice Hall, ISBN: 0-13-334186-0.

[3] Hazas, M., and Ward, A. (2002). A novel broadband ultrasonic location system. In International Conference on Ubiquitous Computing, 264–280.

[4] Hightower, J., and Borriello, G. (2001). Location systems for ubiquitous computing. Computer 34, 8 (Aug.), 57–66.

[5] Huber, D., M., Runstein, R., E. (2005) Modern Recording Techniques. Sixth edition, Focal Press. ISBN: 0240806255.

[6] Olson, E., Leonard, J., and Teller, S. (2006). Robust range only beacon localization. Journal of Oceanic Engineering 31, 4 (Oct.), 949–958.

[7] Priyantha, N., Chakraborty, A., and Balakrishnan, H. (2009). The cricket location-support system. In International Conference on Mobile Computing and Networking, 32–43.

[8] Vallidis, N. M. (2002). WHISPER: a spread spectrum approach to occlusion in acoustic tracking. PhD thesis, University of North Carolina at Chapel Hill.

[9] Randell, C., and Muller, H. L. (2001). Low cost indoor positioning system. In International Conference on Ubiquitous Computing, 42–48.

[10] Volná, E., Jarušek, R., Kotyrba, M., Janošek, M. and Kocian, V. (2011). Data extraction from sound waves towards neural network training set. In R. Matoušek (ed.): Proceedings of the 17th International Conference on Soft Computing, Mendel 2011, Brno, Czech Republic, pp. 177-184. ISBN 978-80-214-4302-0, ISSN 1803-3814.

[11] Volná, E., Jarušek, R., Kotyrba, M. and Rucký, D. (2013) „Dynamical Motion Capture System Involving via Neural Networks". In Banerjee, S. and Erçetin, Ş.Ş. (eds.) The proceedings of Symposium of Chaos, Complexity and Leadership, ICCLS2012 (Springer Complexity series) – in press.

[12] Ward, A., Jones, A., and Hopper, A. (1997). A new location technique for the active office. Personal Communications 4, 5 (Oct.), 42–47.

[13] Welch, G., and Foxlin, E. (2002). Motion tracking: no silver bullet, but a respectable arsenal. Computer Graphics and Applications 22, 6 (Nov./Dec.), 24–38.