Neural Computation of Direction-Of-Arrival of Sound

JACEK CZERNIAWSKI, ANDRZEJ CZYZEWSKI, RAFAL KROLIKOWSKI
Sound & Vision Engineering Department
Technical University of Gdansk
Narutowicza 11/12; 80-952 Gdansk
POLAND

Abstract: - One of issues related to videoconferencing is addressed in the paper, namely - the problem of sound source localization. A new neural approach for estimation direction of arrival (DOA) is presented and discussed therein. The introduced algorithms are based on various types and structures of neural networks, including feedforward and recurrent ones. Considerations of the proposed DOA estimation are supported by some of the results of numerous experiments, carried out using audio material recorded in an anechoic chamber and acoustically adopted room. The results are briefly discussed.

Key-Words: - sound source localization, feedforward & recurrent neural networks

1 Introduction

Localization of sound sources is vital in contemporary tele- and videoconferencing systems as well as hand-free communication sets. In general, people have difficulty in understanding speech with ambient noise, high reverberation or with many concurrent speakers. It is due to the fact that audio signals coming from various sources not only interfere with the target signal but also can obscure it. In order to overcome this problem, the unwanted & parasite signals are attenuated by means of spatial filtration, performed by beamforming techniques [13] [16]. These techniques can be improved when the position of a speaker is known in the 3-D space. However, practically the exact position cannot be obtained and in such a case, the direction of arrival (DOA) for the speaker's acoustic waves is only estimated. Nevertheless, the DOA estimation can considerably increase the efficiency of the acquisition of the audio, since it can reduce influences of other sources on the target signal, improves the signal-to-noise ratio and in result - the effectiveness of relevant noise reduction- & dereverberation algorithms. This is the reason why the localization of sound sources plays important role in the systems mentioned above.

Under ideal conditions the DOA estimation is a straightforward and deterministic problem. However under real conditions there can occur various distortions as background noise and/or reverberated signals interfering with the target one, which makes the problem more complex and non-deterministic. This is why the automatic identification of directionality of sound sources is still unsolved and hence a number of various methods have been proposed. Most of them are based on the estimation of the time delay between signals coming from an array of microphones applying correlation techniques [2], adaptive filtration [3] or computation of relevant eigenvalue vectors and matrices [17]. In turn, in the case of tracking or localization of many sources, the Bayesian-based methods are exploited [19]. More details can be found in the abundant literature on this topic [11] [12] [14].

Despite of the all methods above, in recent years another approach to the DOA estimation problem has been established. In compliance with this approach DOA is estimated by means of signal processing by artificial neural networks (ANNs) [1] [6]. In general, ANNs are quite attractive to be applied in signal processing, since they offer as capabilities as: non-linear approximation [20], dealing with uncertain information [5], modeling of time series [8] and mapping of a complex process dynamics [7]. Although the most common feedforward structures, often referred to as multilayer perceptrons (MLPs), can be employed for various tasks, the latter two features are especially peculiar to recurrent neural networks (RNNs).

Therefore owing to the afore-cited capabilities of ANNs, the authors' proposal of the sound source localization method apply such networks. However, prior to the neural processing, audio signals are parametrized in order to make the DOA estimation more reliable. The parametrization can take place both in the time- and spectral domain. The introduced parameters are originated from psychoacoustics, according to which perception of sound directivity by the human binaural system is based on the following two fundamental entities [10]:

- *Interaural Level Difference*: difference of intensities of waveforms in the left and right ear.
- *Interaural Time Difference*: difference of arrival times of relevant waveforms in the both ears, which is equivalent to a phase difference of these waveforms.

In turn as far as the DOA estimation is concerned, for comparison purposes both types of neural networks are considered, namely: MLPs and RNNs. The rationale for using RNNs is the fact that there are some temporal relationships between signals received by an array of microphones, and thereby these signals can be concerned as time series.

In order to verify the authors' proposal of the sound source localization method, numerous experiments were organized and carried out. Some of the obtained results are included in this paper, as well as a description of the method with a brief discussion is presented therein.

2 Sound Source Localization Method

The general scheme of the method for sound source localization is presented in the figure below (Fig. 1).

Fig. **1**. Scheme of the sound source localization method

In the *Acquisition of Audio* module acoustic waves are received by an array of *L* microphones, and next are converted into digital representation. At the output, a multichannel (*L*-channel) digital signal is obtained. Since neural processing of time-domain audio samples is highly inefficient, some parameters reflecting spatial directivity of the sound are extracted form the signal, which takes place in the *Parametrisation of Audio Signals* block. Outputs of this block - vectors of parameters - are fed to the input of *Neural Estimator* which estimates the direction (angle value) of the relevant sound source.

2.1 Acquisition of Audio

The acquisition of audio is performed with the use of a circular array of omnidirectional microphones. The array consists of L = 8 electret microphones which are uniformly set on the circumference of a 15 cm radius rim. Subsequently, 8 mono tracks are recorded simultaneously and put through the A/D conversion. Thus, the multichannel signal can be described by the following equations:

$$\begin{cases} x_1(t) = \alpha_1 \cdot h_1(t) * s(t) + n_1(t) \\ \vdots \\ x_i(t) = \alpha_i \cdot h_i(t) * s(t - \tau_i) + n_i(t) \\ \vdots \\ x_L(t) = \alpha_L \cdot h_L(t) * s(t - \tau_L) + n_L(t) \end{cases}, \quad (1)$$

where $x_i(t)$ denotes a signal received by the *i*-th microphone and delayed by τ_i with respect to the first (reference) microphone, whereas $s(t)$ stands for a source signal which is attenuated in the same microphone by α_i and distorted by an ambient noise $n_i(t)$ received by this microphone (formally, $h_i(t)$ is referred to as an impulse response of the reverberant channel).

2.2 Parametrization of Audio

As was already mentioned, the introduced parameters are computed either in the time- or spectral domain. The first one is related to a cross-correlation between signals in two different channels, whereas the latter one is based on magnitude & phase relations between respective spectral bins in two different channels.

2.2.1 Time-domain Parameters

The essential information on the temporal relationship between the signals: $x_i(m)$ in the *i*-th channel and $x_j(m)$ in the *j*-th one can be obtained by taking into account the maximum values of the correlation coefficient $\rho_{ij}(\Delta)$ and the respective lag Δ_{ij} between these signals as follows:

$$\rho_{ij} = \rho_{ij}^{\max}(\Delta) = \max_{M+1,\ldots,M-1} \left\{ \rho_{ij}(\Delta) \right\}, \quad (2)$$

$$\Delta_{ij} = \arg\left\{ \rho_{ij}^{\max}(\Delta) \right\}, \quad (3)$$

which are computed in the *M*-point analysis window. Thus, the vector of parameters is composed of the pairs (ρ_{ij}, Δ_{ij}) for a given combination of channels. It seems reasonable to consider only combinations between opposite microphones, since the correlation coefficients for the respective signals are sufficiently distinctive. Hence, this approach results in 8-element vector of parameters.

2.2.2 Spectral-domain Parameters

The spectral-domain parameters reflect time- and level differences between signals in two different channels, which refers to the psychoacoustical fundamentals of perception of sound directivity.

Considering a pair of the *i*-th and the *j*-th channel, the introduced parameters M_{ij}^k and A_{ij}^k can be formulated as follows:

$$M_{ij}^k = \frac{\min\left(\left|Ch_i^k\right|, \left|Ch_j^k\right|\right)}{\max\left(\left|Ch_i^k\right|, \left|Ch_j^k\right|\right)}, \tag{4}$$

$$A_{ij}^k = \angle Ch_i^k - \angle Ch_j^k, \tag{5}$$

where $\left|Ch_i^k\right|$ and $\angle Ch_i^k$ represents the magnitude and the phase of the *k*-th spectral bin of a signal in the *i*-th channel.

In this case, unlikely to the time-domain parametrization, the following combinations as sets of parameters for -channel signals can be considered:

- *type **A***: all mutual combinations of channels, which yields 56 parameters per a bin.
- *type **B***: combination of opposite channels, which yields 8 parameters per a bin.

On account of the fact that the above parameters are to be fed to a neural network, they are grouped into input vectors. The following three types of such vectors can be considered:

- *type **V*1**: all spectral bins are included in a vector.
- *type **V*2**: an input vector consists of parameters for a single bin and the additional information on the bin's frequency.
- *type **V*3**: an input vector consists only of parameters for a single bin. In this case, a neural network assumes a structure of a modular network where a separate neural subnet is dedicated for each spectral bin. The final neural decision is made on the basis of the maximum outputs of all subnetworks.

The choice of the parameters was made considering the following issues: the size of an input vector, a number of training vectors (the size of a training set) and the storage complexity. A detailed discussion on the selection of the parameters is included in one of the recent authors' paper [4], and according to its results & conclusions the following sets of parameters were chosen for the experimental verification:

- Vector type ***V*1**, parameters type ***A***, the size of an analysis frame $N = 512$.
- Vector type ***V*3**, parameters type ***A***, all sizes of an analysis frame ($N = 512$, $N = 1024$, $N = 2048$).
- Vector type ***V*3**, parameters type ***B***, all sizes of an analysis frame ($N = 512$, $N = 1024$, $N = 2048$).

It should be stressed that such a parametrization is requires vast amount of a computer's memory.

2.3 Neural Estimation

To a certain extent, the forms of parameters impose a respective neural structure. As far as the spectral-domain parameters are considered, only two types of vectors of parameters: ***V*1** and ***V*3** can be taken into account. Moreover, it can be noticed that these vectors consist of a great number of elements. Too great so that RNNs could be applied in practice, since the computational & memory complexity is of the order $O(N^3)$ for the *N*-dimensional vectors. Therefore only feedforward structures can be used, yet what makes the difference is that for vectors ***V*1** standard MLPs are applied, and vectors ***V*3** can be processed only with the use of modular networks.

In turn, the time-domain parameters on the one hand are not so much numerous, and on the another - reflect some temporal relationships. Hence, this type of vectors are processed by RNNs.

2.3.1 Employment of Time-domain Parameters

In a given moment of time, a multichannel signal is described by an eight element vector, and this vector becomes an input vector fed to a RNN of the Elman type [8]. Hence the input layer consists of 8+1 neurons (incl. the dummy one), whereas the number of units in the output layer is determined by the number of expected or desired directions. In turn, the size of the hidden- and the context layer is variable, and set experimentally while testing. As regards the neuron's activation function, sigmoidal functions are provided in order to obey the requirements of a gradient-type training algorithm.

2.3.2 Employment of Spectral-domain Parameters

As regards MLPs, processing vectors of the type ***V*1** (parameters for all spectral bins), the size of the input- & output layer is fixed and determined by the dimension of the vector ***V*1** and the number of expected angles, respectively. Only the number of hidden neurons is variable and can be altered during experiments.

In turn, the proposed neural structure for processing vectors of the type ***V*3** is more complex. Namely, a separate MPL is associated to every spectral bin, and is trained to recognize parameters related only to this component, which is presented in Fig. 2.

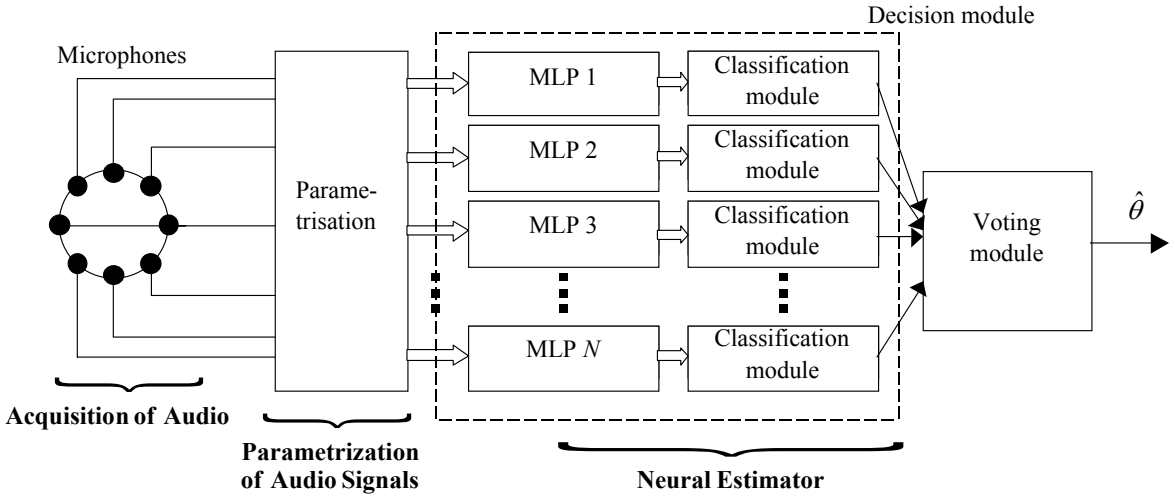

Fig. **2**. Scheme of the modular neural network

Given N different bins in the frequency domain, the total number of MLPs employed is equal to N. The outputs of particular MLPs are fed to *Classification Module*, where the DOA classification for a single spectral component is done. In this module, the decision is made with the use of a hard-defined function, for which the threshold value is set to 0.5. And last but not least - basing on all partial classifications, *Voting Module* estimates the input audio's DOA by means of majority voting. In result, at its output the estimate of the direction $\hat{\theta}$ is obtained.

3 Experimental Estimation of DOA

The recordings were made in an anechoic- and acoustically adopted chamber, where the circular array was fixed 1.58 m from the floor, and there was one male speaker, distanced 1.5 m from the array. The speaker read non-sense syllables (logatoms) from the consecutive spots differing in 5^o, what resulted in 72 eight-track recordings (~ 55 s each). 8 mono tracks (16 bit/sample, 48 kHz) were recorded simultaneously.

For the purposes of this paper, some results related to the sound directivity from -45^o to $+45^o$ and from -30^o to $+30^o$ every 15^o are presented.

3.1 Experiments for Time-domain Parameters Processed by RNNs

In the case of RNNs, only the number of hidden neurons is variable, and hence for purposes of the experiments, the number was arbitrary set to 10. As a neuron's activation function, the unipolar continuous one was chosen. The number of the vectors per class was equal to 100, which yielded totally 500 vectors for a training- and testing phase. The training and testing vectors were different and selected randomly. Besides, 3 values for the length of time sequences T were tested: $T \in (1,2,3)$. In order to obtain statistically valid results, computations were repeated 10 times per a given survey. The RNNs were train with the use of an algorithm based on the standard one by Wiliams & Zipser [18].

Tab.**1**. Localization results vs. particular directions

θ	Scores [%]			
	max			avg.
	T = 1	T = 2	T = 3	
-30^o	94.6	96.1	94.1	94.9
-15^o	93.8	94.7	95.0	94.5
0^o	95.6	96.4	92.1	94.7
15^o	91.3	94.4	94.3	93.3
30^o	96.0	95.9	96.2	96.0

The results of the experiments are assembled in Tab. 1, where the maximum- and average classification scores are given for various lengths of times series and sound source directions.

3.2 Experiments for Spectral-domain Parameters Processed by MLPs

Since for MLPs the size of the hidden layer could alter, thus the number of hidden neurons changed in the arbitrary range: from 20 to 50. Depending on the size of the analysis window (N) and type of vectors of parameters, the training-to-test vectors ratio assumed: 1024 / 446, 515 / 221, 252 / 108. As for RNNs, the unipolar continuous function was chosen as a neuron's activation one. As regards training, the adaptation of MLPs' weight values was achieved using (alternatively) one of the following standard training algorithms: the Fahlman's QuickPROP [9] and the **R**esilient **PROP**agation (RPROP) [15].

In Tab. 2-4 there are included results: the maximum- and the average score vs. particular directions for a number of different surveys:

- parameter vectors of the type ***V*1** are processed by MLPs (Tab. 2).
- vectors type ***V*3**, including parameters type ***A***, are processed by a modular network (Tab. 3).
- vectors type ***V*3**, including parameters type ***B***, are processed by a modular network (Tab. 4).

Tab. **2**. Localization results vs. directions: MLPs in use

θ	Scores [%]	
	max	avg.
-45°	86	84.7
-30°	83	81.0
-15°	84	82.3
0°	86	83.3
15°	82	79.7
30°	84	80.7
45°	83	81.7

Tab. **3**. Localization results vs. directions: parameters type ***A*** with modular nets in use

θ	Scores [%]	
	max	avg.
-45°	92	88.0
-30°	89	87.3
-15°	90	88.3
0°	90	89.0
15°	87	85.2
30°	86	85.2
45°	88	87.3

Tab. **4**. Localization results vs. directions: parameters type ***B*** with modular nets in use

θ	Scores [%]	
	max	avg.
-45°	85	82.8
-30°	84	81.6
-15°	83	80.8
0°	83	81.2
15°	83	81.3
30°	84	81.6
45°	83	80.7

During the experiments, the increment of the efficiency could be observed dependently on the number of hidden neurons as follows: for 35, 40 and 45, the total effectiveness achieved 83.4 %, 88.9 % and 89.5 %, respectively. Moreover, it could be noticed that a sort of audio signals influenced the scoring, namely: the best efficiency was observed for loud excerptions, whereas the worst one - for noisy ones. In turn, as regards single MLPs in the modular networks, their effectiveness was rather poor (approx. 40-70 %), however good total classification ratios were obtained thanks to *Classification*- and *Voting Modules*.

4 Conclusions

In the paper a brief discussion on the estimation of DOA by means of neural networks was presented. The discussion encompassed the analysis of digitized audio signals both in the time- & the frequency domain, and as the result - the introduced parametrization of the audio is based on either temporal relationships between the signals in different channels or some principles of psychoacoustics. As regards neural processing, the proposed estimators are based on feedforward and recurrent structures.

The obtained results suggest that correlation parameters computed in the time-domain and processed by RNNs are the most efficient. On the other hand, RNNs are characterized by considerable computational & memory complexity, and furthermore - it is more difficult for training algorithms to converge than in the case of the standard MLPs. In turn, the employment of MLPs for processing of the spectral-domain parameters can improve the convergence of training algorithms. However, the need of the time-to-frequency transformation, computation over numerous MLPs and the great size of vectors make the DOA estimation of this type very time-consuming.

Nevertheless, the results of the experiments are quite interesting. Therefore the neural approach for the DOA estimation could be an effective alternative for standard DOA methods.

Acknowledgements

The research is sponsored by: the Foundation for Polish Science and the State Committee for Scientific Research, Warsaw, Poland. Grant No. 8 T11D 002 18.

References:

[1] G. Arslan, F. Sakarya, A Unified Neural-Network-Based Speaker Localization Technique, *IEEE Trans. on Neural Networks*, Vol. 11, No. 4, July 2000, pp. 997-1002.

[2] M. Brandstein, A Pitch-Based Approach to Time-Delay Estimation of Reverberant Speech, *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 1997.

[3] S. Chern, S. Lin, An Adaptive Time Delay Estimation with Direct Computation Formula, *J. of Acoustical Society of America*, Vol. 96, No. 2, February 1994, pp. 811-820.

[4] A. Czyzewski, B. Kostek, R. Krolikowski, Neural Networks Applied to Sound Source Localization, *Proc. of the* 110^{th} *Audio Engineering Society Convention*, Amsterdam, Holland, 2001, Preprint No. 5375.

[5] A. Czyzewski, R. Krolikowski, Neuro-Rough Control of Masking Thresholds for Audio Signal Enhancement, *J. of Neurocomputing*, Vol. 36, No. 1-4, February 2001, pp. 5-27.

[6] M. Datum, F. Palmieri, A. Moiseff, An Artificial Neural Network for Sound Localization Using Binaural Cues, *J. of Acoustical Society of America*, Vol. 100, No. 1, July 1996, pp. 372-383.

[7] S. Day, M. Davenport, Continuous-Time Temporal Back-Propagation with Adaptable Time Delays, *IEEE Trans. on Neural Networks*, Vol. 4, No. 2, March 1993, pp. 348-354.

[8] J. Elman, Finding Structure in Time, *Cognitive Science*, Vol. 14, 1990, pp. 179-211.

[9] S. Fahlman, An Empirical Study of Learning Speed in Back-Propagation Networks, *Technical Report of Carnegie Mellon University in Pittsburgh* (*USA*), No. CMU-CS-88-162, September 1988.

[10] W. Hartmann, How We Localize Sound, *Physics Today*, Vol. 11, November 1999, pp. 24-29.

[11] G. Jacovitti, G. Scarano, Discrete Time Techniques for Time Delay Estimation, *IEEE Trans. on Signal Processing*, Vol. 41, No. 2, February 1993, pp. 525-533.

[12] F. Khalil, J. Lullien, A. Gilloire, Microphone Array for Sound Pickup in Teleconference Systems, *J. of Audio Engineering Society*, Vol. 42, No. 9, September 1994, pp. 691-700.

[13] H. Krim, M. Viberg, Two Decades of Array Signal Processing Research: The Parametric Approach, *IEEE Signal Processing Magazine*, Vol. 13, No. 2, April 1996, pp. 67-94.

[14] Y. Mahieux, G. le Tourneur, A., Saliou, A Microphone Array for Multimedia Workstations, *J. of Audio Engineering Society*, Vol. 44, No. 5, May 1996, pp. 365-372.

[15] M. Riedmiller, Advanced Supervised Learning in Multi-layered Perceptrons - from Backpropagation to Adaptive Learning Algorithm, *Intl. J. of Computer Standards and Interfaces*, *Special Issue on Neural Networks*, Vol. 5, 1994.

[16] B. Veen, K. Burckley, Beamforming: A Versatile Approach to Spatial Filtering, *IEEE Signal Processing Magazine*, Vol. 5, No. 2, April 1998, pp. 4-24.

[17] S. Visuri, H. Oja, V. Koivunen, Subspace-Based Direction-of-Arrival Estimation Using Nonparametric Statistics, *IEEE Trans. on Signal Processing*, Vol. 49, No. 9, September 2001, pp. 2060-2073.

[18] R. Wiliams, D. Zipser, A Learning Algorithm for Continually Running Fully Recurrent Neural Networks, *Neural Computation*, Vol. 1, 1989, pp. 270-280.

[19] I. Ziskind, M. Wax, Maximum Likelihood Localization of Multiple Sources by Alternating Projection, *IEEE Trans. on Acoustics*, *Speech and Signal Processing*, Vol. 36, No. 10, October 1988, pp. 1553-1560.

[20] J. Zurada, *Introduction to Artificial Neural Networks*, West Publishing Company, St. Paul New 1992.