Classification of Spatial Audio Location and Content Using Convolutional Neural Networks

Toni Hirvonen
Dolby Laboratories, Inc.

Audio Engineering Society

Convention Paper

Presented at the 138th Convention
2015 May 7–10 Warsaw, Poland

Classification of Spatial Audio Location and Content Using Convolutional Neural Networks

Toni Hirvonen[1]

[1]*Dolby Laboratories. Stockholm, Sweden*

Correspondence should be addressed to Toni Hirvonen (`toni.hirvonen@dolby.com`)

ABSTRACT

This paper investigates the use of Convolutional Neural Networks for spatial audio classification. In contrast to traditional methods that use hand-engineered features and algorithms, we show that a Convolutional Network in combination with generic preprocessing can give good results, and allows for specialization to challenging conditions. The method can adapt to e.g. different source distances and microphone arrays, as well as estimate both spatial location and audio content type jointly. For example, with typical single-source material in a simulated reverberant room, we can achieve cross-validation accuracy of 94.3% for 40-ms frames across 16 classes (eight spatial directions, content type speech vs. music).

1. INTRODUCTION

Humans have senses that provide information to the perceptual system, allowing us to vary our behaviour depending on the external environment. There has long been a desire to have machines (i.e. algorithms) mimicking or even surpassing the human capabilities in this regard. The first part of the task is not a problem, as sensor technology is being used extensively to gather information in the modern world The main issue is what to do with the overwhelming amount of information in order to benefit from it.

While the nature of human behaviour and the high-level functioning of the perceptual system remain difficult, perhaps even unsolvable problems, there have been notable advances in pattern recognition and machine learning. Many of the impressive results of recent times can be attributed to using old ideas efficiently. Methods that did not quite fulfil their potential few decades ago, have now become standard because of increased computational capacity and more careful implementation details.

”Learning” is indeed a concept that is attractive for modern sensor applications. It is beneficial if the machine can, at least to some extent, function as a "black box". This means that the engineer does not hand-customize every detail of the algorithm to suit the given application, but rather uses more generic methods that self-adapt. Using the black-box learning methods has been in recent times proven the most successful approach in many fields, such as computer vision [1], speech recognition [2, 3], and natural language processing [4]. It seems that hand-crafting approaches can often result in less optimal solutions, especially when heuristics and complicated problems are involved.

In this paper, we are interested in applying the state-of-the-art learning techniques to spatial audio analysis and classification via microphone capture. The microphone is a well-known example of a sensor that is more and more utilized in the modern world. In addition to performing generic audio capture, various microphone array techniques have become popular in academia and consumer products (e.g. Kinect). There they are used for spatial sound field capture [5], spatial sound enhancement, source separation, and estimation of source direction [6, 7].

In typical audio directional analysis systems, handcrafted features and algorithms are predominant. Also, in case of any further classification steps, such as speech recognition, separate classifiers with different features are usually concatenated after the initial spatial analysis. Adapting such traditional methods to specific conditions, e.g. to a different room or microphone array, typically would require a significant human effort.

In the following chapters, we examine an alternative approach: The goal is to demonstrate that microphone array analysis can be made more automatic by utilizing "black-box" methods that have proven to be successful in a variety of other supervised classification tasks. We describe how well-known generic machine learning algorithms can be utilized for directional analysis of the sound field in the same manner as typical specialized algorithms, and also more adaptively if needed. In order to also show that the spatial analysis can be jointly optimized with a typical monaural audio classification task, we simultaneously predict the audio type with speech/music separation in addition to the location. The experiments are done in a simulated room using a virtual microphone array. By merely generating labelled training material, we can optimize a classifier that learns sophisticated features automatically. The general intent of the paper is also to advocate the use of these methods in spatial audio to hopefully inspire further research.

2. BACKGROUND

2.1. Deep Convolutional Neural Networks

During the last decade, there has been a notable renaissance in artificial neural networks for both unsupervised and supervised learning [8]. A significant part this success is attributed to the Deep Learning paradigm, which is understood to have several benefits. Deep networks with large number of parameters require sophisticated algorithms, as well as computational power, to perform well. With the recent development of these aspects, new standards have been set in many classification tasks, as mentioned previously. In this section, we briefly describe the Deep Network aspects that are the most relevant to the present experiments.

One of the first learning machines was the Perceptron. It is a classifier that gives the output y related to the linear combination of the input features $x \in \mathbb{R}^n$ (plus an optional bias term):

$$y = sgn(w^T x + b) \tag{1}$$

The weight vector $w \in \mathbb{R}^n$ is learned in a supervised manner. However, if the input is not linearly separable, no perfect classification can be achieved. This poses high demands on the designer in how to best construct the features x. Many practical applications of machine learning today use "glorified" linear classifiers or template matching [8], so this issue is relevant. As mentioned we are interested in moving toward increased learning capacity and more generic data preprocessing.

In order to model non-linear relationships, we can combine Perceptron-like neuron units into multiple layers after the input layer. Each layer has multiple neurons, and in the simplest case there is full connectivity between the layers (see Fig. 1a). When terminated by a specific output layer such as the softmax function, a neural network is formed. If there is just one hidden layer between the input and output layer, the network is called shallow. A Deep Neural

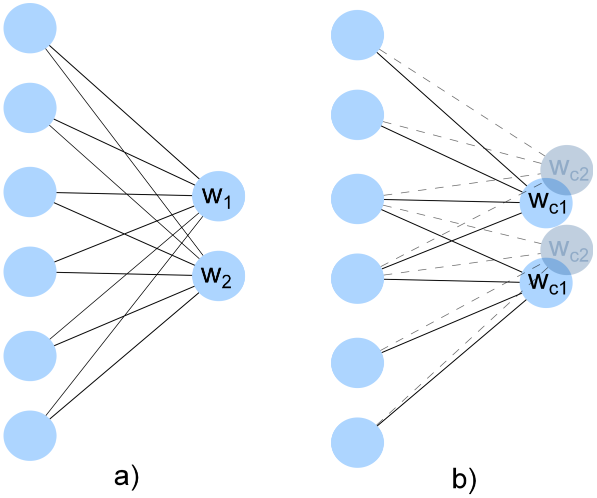

Fig. 1: Example network layers. For a fully connected layer a), there are separate input weights $w_1 \in \mathbb{R}^6$ and $w_2 \in \mathbb{R}^6$ for both neurons. In the convolutional layer b), there is only local connectivity. The weights $w_c \in \mathbb{R}^4$ are also shared between neurons reducing the number of learned parameters. Typically, b) has multiple filters formed by the neurons with the same weights.

Network is achieved by having more than one hidden layer. Intuitively, we can have higher level abstractions with more layers. The primary advantage is that deep nets can compactly represent a significantly larger set of functions than shallow networks [8].

Multilayer nets can be trained in a supervised manner by the backpropagation algorithm. It trains a discriminative model, i.e. the network models the conditional distribution $P(y|x)$ between the input and output. Typically used stochastic gradient descent updates the network weights W by a linear combination of the negative gradient of the loss function $L(W)$, and the previous weight update V_t:

$$V_{t+1} = \mu V_t - \alpha \frac{\partial L(W_t)}{\partial W} \tag{2}$$

$$W_{t+1} = W_t + V_{t+1} \tag{3}$$

where α indicates learning rate and μ the momentum aka. the influence of the previous update. The success of backpropagation is sensitive to the details in the data preprocessing, the network, and the learning parameters [8, 9].

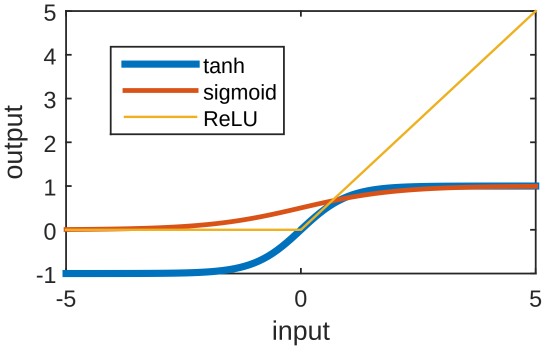

Fig. 2: Commonly used neuron activation functions.

One of these details is the non-linear activation function $f(x)$. It is used for processing the output of each neuron. Lacking these functions, the multilayer net would reduce to a linear function. The Perceptron used simple sign-function, and functions like sigmoid or hyperbolic tangent were popular in the past. However, the Rectified Linear Unit (ReLU) is nowadays almost the de facto choice because of it's low-complexity (both in forward pass and backward gradient), and the fact that it avoids the vanishing gradient-issue in backpropagation [9, 10]. The latter is because the ReLU output is not limited in magnitude. As seen in Fig. 2, the ReLU activation function is given by:

$$f(x) = \begin{cases} 0 & : x < 0 \\ x & : x \geq 0 \end{cases} \tag{4}$$

An issue with the deep fully connected discriminate networks is the sheer number weight parameters w to be optimized. If we can reduce the number of parameters, classification robustness and training data requirements generally benefit. A convolutional layer [11] is a way of achieving this, while still allowing for expressive deep architectures. Fig. 1b shows the difference to a fully connected layer. The neuron connectivity is limited to be local instead of full, and the weights are shared across neurons. This can be interpreted as a convolution operation with a kernel w

that is learned, hence the name of the convolutional layer.

Convolutional networks were initially utilized, and continue to be state-of-art, for image object recognition, since they work well against image translation. However, they are also increasingly utilized in other applications. In audio research, generative convolutional nets have been used for speech recognition and other audio classification tasks [12]. We are not aware of any specific use of discriminative convolutional nets for spatial audio location estimation.

2.2. Source Localization

Acoustic source localization via microphone array signal analysis is a widely-researched topic [6]. Numerous applications include e.g. speaker tracking [13], robotics, [14], hearing aids [15], and security applications [16]. Often acoustic source localization is thought of as the problem of determining the Direction of Arrival (DOA). For the methods studied in this paper, the distinction between DOA estimation and general localization is however relevant, as is be discussed in Section 4.1.

DOA estimation algorithms can be based on various approaches. Early research for single sources relied on analysing the time difference [17]. Subspace-based methods like MUSIC [14] then became popular in general array processing and can be utilized for simultaneous multiple sources. As an example of the current state-of-art, the method of [18, 7] has a relatively low computational cost and relaxed assumptions while providing good accuracy for a circular array. It is used in this paper as a benchmark.

The benchmark method relies on a simple anechoic signal propagation model. The DOA estimation relies on correlation analysis, and the known propagation delays between the M different microphones. The estimated azimuth angle θ_ω per each frequency ω is calculated from the Circular Integrated Cross Spectrum (CICS):

$$\theta_\omega = \arg\max_\phi |\mathrm{CICS}^{(\omega)}(\phi)| \tag{5}$$

where

$$\mathrm{CICS}^{(\omega)}(\phi) = \sum_{i=1}^{M} e^{-j\omega\tau_{m_i \to m_1}(\phi)} G_{m_i m_{m+1}}(\omega) \tag{6}$$

$G_{m_i m_{m+1}}(\omega)$ indicates the cross-power spectrum between microphones m_i and m_{m+1} divided by it's absolute value. The difference in the relative delay between the microphone pairs $m_1 m_2$ and $m_i m_{i+1}$, given the input signal angle ϕ, is indicated by $\tau_{m_i \to m_1}(\phi)$.

For the final DOA, we use the mode (most common value) of the θ_ω over frequency, and do not consider further histogram analysis [7]; we only analyse single frames and do not have multiple simultaneous sound sources in the experiments. While some reflections are present in our simulations, the method works well in practice. However, relying on the propagation delay does not distinguish sources at different distances, and that would require non-trivial modifications to the array signal model. Thus, typical algorithms focus only on DOA estimation.

2.3. Speech/Music Classification

Classifying audio segments either as speech or music is a relatively simple problem with only two classes, and there is an abundance of labelled data for training and testing. As an example industrial application, general purpose audio codecs like the recent USAC [19] have utilized such classifiers to switch between speech- and general audio modes with good accuracy.

Speech/music classification has predominantly been researched from an audio engineering perspective. Typically, an audio expert uses knowledge-based methods to hand-craft features that intuitively can discriminate between music and speech. The benefit of this approach is good performance while using relatively few features. A widely cited method using 13 features and Gaussian Mixture Models is given in [20]. More recent research has been much focused on refining the complexity. It was shown in [21] that only two features can be sufficient. However, at least some of the essential traditional features, like zero crossing variance, are calculated over long audio segments.

In contrast to the more traditional approaches, both supervised [22] and unsupervised [23] automatic feature learning methods were recently utilized for this problem. It was shown that even for short frames standalone, these methods can produce close to comparable performance to the methods with long lookback and/or lookahead [22]. This gives indication that black-box learning can prove successful with this and related problems. The question examined in the following is that can we combine the problem jointly with other classification tasks.

3. METHODS

3.1. Research Question

The research question of this paper is: Can we use CNNs for spatial audio analysis and classification? While CNN capabilities are well known, the processing chain that is suitable for all the tasks we are interested in, is not obvious. A generic, flexible pre-processing would be desirable in order to adapt automatically to different situations. We examine this hypothesis by jointly classifying both spatial location and audio type, taking source distance into account, and varying the microphone array type.

The intent of the paper is not to set a classification benchmark. It is difficult to find or construct a de-facto microphone array material database, whose results would be generalizable to all important conditions. Rather, we demonstrate a chain of machine learning techniques, which can be successfully used for such problems. The classification accuracy numbers are mainly used for internal comparisons between test variations.

3.2. Data Preparation, Room Simulation, and Pre-Processing

The database used for CNN training and evaluation included a separate music (many styles, 242 unique files, total 9.4 hours) and speech set (many languages and talkers, 282 files, total 6.6 hours). The datasets attempt to capture a large range of realistic situations that a typical classifier operating for multimedia content would encounter. To get the spatial training and testing material, we use a virtual 2-D microphone array with eight microphones in the centre of a room.

The microphone impulse responses were simulated in a typical reverberant room with T60=0.2 second and dimensions 6 x 4 x 3 meters using a freely available room modelling software [24]. We used two different arrays: 1) the default circular array with eight microphones at [0,45,90,135,180,225,270,315]° and at radius 0.05 m from the array centre; 2) an experimental mixed radius array that was otherwise like circular, except the radius of microphones at [45,135,225,315]° was increased to 0.1 m. Fig. 3 illustrates the locations of the microphones in the two arrays. The microphones are in the same plane in terms of height, i.e. elevation is not considered in this study. The microphone responses were convolved with random segments from the music/speech database according to what azimuth and content type was required for a given frame.

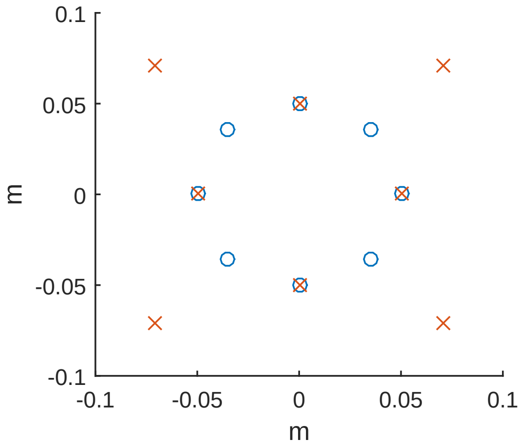

Fig. 3: The locations of the microphones in the two arrays that were studied, as seen from above. Circular array microphone is indicated by (○), and the mixed radius array microphone by (x). Four microphones of the two configurations overlap.

The raw input data for the network training and testing were spectrograms using a 40 ms frame size. As first step, we calculate the compressed log signal power in logarithmically spaced frequency bands from the signal Short-Time Fourier Transform X. Comparing logarithmic banding to bin grouping based on Equivalent Rectangular Bandwidth did not seem to affect results notably. The number of frequency bands can be varied according to complexity requirements, here it was 128. For each band of index i containing bins b_i, the value of the final spectrogram S was:

$$S(i) = 10\log_{10}(\sum_{b_i}(|X(b_i)|^2)/\text{length}(b_i) + 1) \quad (7)$$

The eight microphone spectrograms are concate-

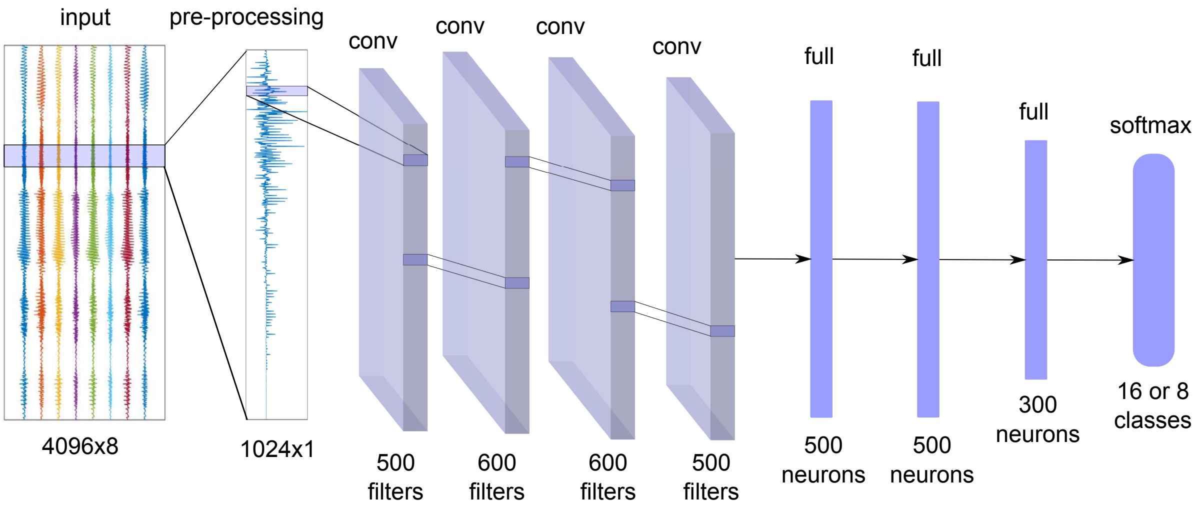

Fig. 4: The CNN architecture for the experiments. A convolutional layer is indicated by "conv", and full-connectivity layer by "full".

nated to form a single 1024-dimensional vector of spectrogram values per frame.

Training of the CNN proceeded as follows: The gathered band vector data is normalized in terms of mean and standard deviation across spectrogram bands, and whitened (a transformation resulting in an identity covariance matrix) using Zero-Phase Component Analysis (ZCA) [25]. This is a typical machine-learning step, and here it removes the redundant information between spectrogram bands and equalizes the components [9]. The cross-validation testing step involved a similar gathering of segment spectrograms, which are ZCA whitened and normalized using the training data covariance, mean, and standard deviation.

3.3. Network Structure

Fig. 4 outlines the CNN architecture for the experiments in this paper. We use a total of seven layers with learned parameters, and a final softmax layer for classification. The first four layers have convolutional connectivity, as discussed in Section 2.1. A seemingly useful strategy for lowering the number of network parameters was to use short strides for the convolutional layers. In this study, the stride was 4, with an overlap of 2. Pooling layers typical to vision research are omitted here. They do not seem useful for audio, as positional pattern shifts do not occur in the whitened spectrograms as in visual images.

The last three layers have full connectivity. We use standard backpropagation with stochastic gradient descent to train the CNN. The learning hyperparameters were: learning rate $\alpha = 0.01$ and momentum $\mu = 0.9$. Throughout the training, α was decreased sequentially by multiplying it by 0.1 every 160000 training samples. The batch size for the weight updates in training was 16 training samples.

In each separate condition, we use 1120000 training samples witch were always randomized segments from the database. Training consisted of one pass thought this dataset. The testing phase utilized a dataset of 160000 different random samples. Each dataset had an equal number of each class instance in a given case. The network computations were realized with Caffe [26], which is a widely used open-source CNN implementation. Caffe utilizes graphical processing units for computations making it possible to train large networks in a reasonable time.

4. RESULTS

Below, we discuss results obtained by training the CNN architecture described in Section 3.3. Various training and testing conditions were simulated

	Spatial 8 Classes	Spatial, Type 16 Classes
Circular array (CA)	97.9	93.7
CA, distant sources	96.3	92.2
Mixed radius array	97.4	94.3

Table 1: Cross-validation percent correct results for the test conditions described in Sections 4.1.-3.

in a room specified in Section 3.2. Table 1 summarizes the classification performances as percent correct values for 40 ms frames. Analysing classifications over multiple frames with averaging, or more sophisticated methods, naturally gives even better performances, but is not detailed here.

4.1. Source Localization vs. DOA Estimation

We first consider classifying sources based solely on their spatial location, disregarding the type of audio. In order to compare with previous DOA studies, we simulated eight sources at [0,45,90,135,180,225,270,315]° and at radiuses 1.5 m from the room centre, where the microphone array was located. The sources and the array were both at 1.5 m height. Only one source was active for given training or testing sample, and thus the eight spatial classes are mutually exclusive.

With the circular microphone array (Fig. 3), we get a cross-validation accuracy of 97.9%, i.e. 2.1% of the random 40-ms frames were misclassified as originating from a false direction. To compare this with the current state-of-art DOA estimation, we implemented the method described in Section 2.2 with limiting the possible source azimuth angles ϕ only to the eight source directions. This gives a better accuracy of 98.6% for the same array and test set.

In order to understand why this happens, it should be remembered that the generic data preprocessing used here actually throws away information. By taking only the absolute value of the microphone signal spectrograms and combining over frequency bands, we seemingly disregard the phase of the STFT. However, we note that some of the signal features after pre-processing change e.g. depending on the amount of reflections. The DOA estimation method, on the other hand, relies on the phase by analysing the cross-power spectra, and knowing the time delays between microphone pairs. Black-box methods like neural networks are attractive when the phenomenon variance can not be well explained by a relatively simple, understandable model. For DOA estimation in a space of low noise and reflections, the simpler delay-model seems to work well.

If we would want to increase the CNN DOA estimation performance, the generic pre-processing should likely be changed to utilize more phase information. This is a topic for future research. However, the accuracy is arguably already quite usable, since we analyse such short frames. Furthermore, the utilized pre-processing allows us to go beyond DOA estimation by classifying generic spatial locations in a specific room. We demonstrate this in another test by having sources with the same DOA but at different distances. We again simulate eight sources with a circular array in the previous room, however now with only four azimuth directions [0,90,180,270]°. There were two sources in each direction, one with distance 1.5 m from the array and the other further away, 0.1 m distance from the wall. This experiment is noted as "distant sources" in Table 1.

The CNN accuracy for location classification is still quite good (96.3%) for discriminating spatial locations in a more generic manner. Sources near the walls have notably different impulse response compared to more central ones with the same DOA, and the CNN learns to exploit this automatically. It is trivial to see that a DOA estimator would fail this test, having the theoretical best-case accuracy of 50%. Hand-engineering a good solution based on e.g. delay and reflection analysis is non-trivial and laboursome. However, despite the advantage of the CNN, the trade-off is that it is learned specific to the array, the sources, and the room. If these aspects change notably, re-training is appropriate.

4.2. Joint Spatial and Content Type Classification

Another benefit of the generic pre-processing is that is allows for simultaneous, joint classification based on various different attributes. To demonstrate this, we train the CNN to discriminate between 16 mutually exclusive classes that were formed by combinations of eight spatial locations, and the audio type (speech vs. music). A circular microphone array with sources at [0,45,90,135,180,225,270,315]° and at radiuses 1.5 m from the room centre was again utilized as a vanilla test setup. The accuracy achieved is 93.7%. This again seems usable for applications.

Once more, hand-engineering features to best perform these tasks jointly would be an intricate task. Nevertheless, it should be noted that it is possible to use independent, hand-crafted DOA estimators and speech/music classifiers simultaneously and simply concatenate their results. Given good room conditions, and assuming 95% speech/music separation accuracy for short frames [22], one may get close to the accuracy of the joint CNN classification. However, difficulties for standard methods arise if we take source distance into account, must adapt to different rooms etc. These problems are alleviated by black-box methods like the CNN. Also, if the two events are, unlike here, mutually dependent so that certain combinations do not occur, the CNN can adapt to that situation given appropriate training data.

4.3. Adaptability and Robustness in Different Conditions

In order to investigate the automatic adaptability of the CNN to different conditions than a standard circular array, we experimented with a mixed radius array detailed in Fig. 3. For spatial localization of equidistant sources to eight classes, the accuracy achieved was 97.4%. This is slightly worse than for the circular array in the same task. However, in the task of joint spatial and content type classification to 16 classes, the mixed array seems to give slightly better results than the circular array (94.3%).

The reason for this interesting difference between the two array performances is not completely clear. Even though the difference is not big, it can be speculated that having more variance in the microphone room responses, as in the mixed array, allows for more perspective for content classification. For localization of equidistant sources, on the other hand, the mixed array might not be so helpful since increasing distances between microphones also introduces temporal aliasing to lower frequencies than in the circular array. Despite the underlying reasons, the test demonstrates that with the black-box methods, it is trivial to experiment with different experimental variables to improve performance in a specific situation.

Since the CNN is trained specifically to a given set of signals, spatial locations, and rooms, one may ask how well an existing network tolerates changes to these conditions. For example, what happens if we have a source in a direction that was not included in the CNN training data? In principle, the training database should include all variations of the samples in a given class that are intended to be similar. One could then train the network for classifying to spatial areas instead of single locations.

A more heuristic option is to just use the existing CNN. Even if the training data does not include certain cases, the system may give a sensible classification if we assume that the similarity between the training samples and the new cases is relative to their "true" distance in some meaningful measure. Investigating the outputs of the softmax layer directly may also give better indication, if e.g. testing a single-source net with multiple simultaneous sources.

As an example robustness experiment, we tested the CNN trained for source directions [0,45,90,135,180,225,270,315]° with displaced sources that deviated from these directions. If the displacement was 1°, the accuracy was diminished by 1.1 percent units, whereas with the 5° displacement, the accuracy fell 18.6 percent units. This was for the circular array.

Interestingly, in the mixed array the accuracy with displaced sources did not diminish as much. With the 1° displaced sources the error increased 0.4 percent units and with 5° displacement test, the result was only 4.7 percent units poorer than with the actual directions the CNN was trained for. While again difficult to explain, this result gives another reason to speculate that with different arrays, the hidden network features can be notably distinct. In this case, this hypothetically manifests as lower sensitivity to DOA in a mixed radius array.

5. SUMMARY AND CONCLUSIONS

In this paper, experiments detailing the use of CNNs to spatial audio analysis were presented. We believe this represents a novel application of a previously established supervised learning method. An important aspect of the paper is to corroborate the use of an audio pre-processing technique that allows for a good classification performance in various even seemingly unrelated tasks. This was not obvious a priori.

The experiments involved a standard DOA type problem using a microphone array, as well as variations in source distances, array configuration, and simultaneous classification of content type and location. It was demonstrated that black-box learning methods like CNN allow for adapting to specific situations where hand-engineering and optimization would be non-trivial and time-consuming. Issues with the CNN method were recognized to be the need for retraining if the specific conditions change much, and the difficulty of explaining the hidden features of the network.

Future research includes preserving phase information in the pre-processing for DOA estimation optimization, analysing multiple frame classifications with averaging, histograms, or even state-based models, as well as analysis of multiple simultaneous sources and noisy real-life conditions.

6. ACKNOWLEDGEMENTS

The author wishes to thank Dr. Per Hedelin for valuable comments and discussion.

7. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, ”ImageNet Classification with Deep Convolutional Neural Networks”, in Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012.

[2] G. Hinton, L Deng, D. Yu, A. Mohamed, N Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, and B. Kingsbury, ”Deep Neural Networks for Acoustic Modeling in Speech Recognition”, IEEE Signal Processing Magazine, Vol 29, No 6, 2012.

[3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, A. Ng ”Deep Speech: Scaling up end-to-end speech recognition”, arXiv:1412.5567v2, 2014.

[4] R. Socher, C. Lin, A. Ng, and C. Manning, ”Parsing Natural Scenes and Natural Language with Recursive Neural Networks”, in Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.

[5] V. Pulkki, ”Spatial sound reproduction with directional audio coding”, Journal of the Audio Engineering Society, vol. 55, no. 6, 2007.

[6] H. Krim and M. Viberg, ”Two decades of array signal processing research - The parametric approach”, IEEE Signal Process. Mag., vol. 13, no. 4, 1996.

[7] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, ”Real-Time Multiple Sound Source Localization and Counting using a Circular Microphone Array”, IEEE Trans. Audio, Speech, and Language Processing, vol. 21, no. 10, 2013.

[8] Y. LeCun and M. Razanto ”Deep Learning Tutorial”, in Proceedings of the 30th International Conference on Machine Learning (ICML), 2013.

[9] Y. LeCun, L. Bottou, G. Orr and K. Mller, ”Efficient BackProp”, in G. Orr and K. Mller. Neural Networks: Tricks of the Trade. Springer. 1998.

[10] X. Glorot, A. Bordes and Y. Bengio, ”Deep sparse rectifier neural networks”, Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), 2011.

[11] K. Fukushima, ”Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. Biological Cybernetics 36, 1980.

[12] H. Lee, P. Pham, Y. Largman and A. Ng, ”Unsupervised feature learning for audio classification using convolutional deep belief networks”, in Advances in Neural Information Processing Systems 22, Curran Associates, Inc., 2009.

[13] D. Bechler, M. Schlosser, and K. Kroschel, ”System for robust 3D speaker tracking using microphone array measurements”, in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2004

[14] S. Argentieri and P. Dans, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics", in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2007.

[15] T. Van den Bogaert, E. Carette, and J. Wouters, "Sound source localization using hearing aids with microphones placed behind-the-ear, in-thecanal, and in-the-pinna", Int. J. Audiol., vol. 50, no. 3, 2011.

[16] J. Sallai, W. Hedgecock, P. Volgyesi, A. Nadas, G. Balogh and A. Ledeczi, "Weapon classification and shooter localization using distributed multichannel acoustic sensors", J. Syst. Arch 57, 2011.

[17] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview", EURASIP J. Appl. Signal Process., 2006.

[18] A. Karbasi and A. Sugiyama, A new DOA estimation method using a circular microphone array, in Proc. Eur. Signal Process. Conf. (EUSIPCO), 2007.

[19] ISO/IEC 23003-3:2012 Information technology – MPEG Audio Technologies – Part 3: Unified Speech and Audio Coding, 2012.

[20] E. Scheier and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", Proc. IEEE Conf. on acoustics, Speech, and Signal Processing (ICASSP), 1997.

[21] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings", IEEE Transactions on Multimedia, Vol. 7, No 1, 2004.

[22] T. Hirvonen, "Speech/Music Classification of Short Audio Segments", IEEE International Symposium on Multimedia (ISM), 2014.

[23] A. Pikrakis and S. Theodoridis, *Speech-Music Discrimination: a Deep Learning Perspective*, Proc. 22nd Europen Signal Processing Conference (EUSIPCO), 2014.

[24] E. Lehmann and A. Johansson, Diffuse reverberation model for efficient image-source simulation of room impulse responses, IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 6, 2010. Available at: http://www.eric-lehmann.com/

[25] Online: http://ufldl.stanford.edu/wiki/index.php/Implementing_PCA/Whitening. Retrieved January 2015.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv:1408.5093, 2014.