Multi-scale Aggregation of Phase Information for Complexity Reduction of CNN Based DOA Estimation

Soumitro Chakrabarty
*International Audio Laboratories Erlangen**
91058 Erlangen, Germany
soumitro.chakrabarty@audiolabs-erlangen.de

Emanuël A.P. Habets
*International Audio Laboratories Erlangen**
91058 Erlangen, Germany
emanuel.habets@audiolabs-erlangen.de

***Abstract*—In a recent work on direction-of-arrival (DOA) estimation of multiple speakers with convolutional neural networks (CNNs), the phase component of short-time Fourier transform (STFT) coefficients of the microphone signal is given as input and small filters are used to learn the phase relations between neighboring microphones. Due to the chosen filter size, $M-1$ convolution layers are required to achieve the best performance for a microphone array with M microphones. For arrays with large number of microphones, this requirement leads to a high computational cost making the method practically infeasible. In this work, we propose to expand the receptive field of the filters to reduce the computational cost of our previously proposed method. To realize this expansion, we use systematic dilations of the filters in each of the convolution layers. Different systematic dilation strategies for a specific microphone array are explored. Experimental analysis of the different strategies, shows that an aggressive expansion strategy results in a considerable reduction in computational cost while a relatively gradual expansion of the receptive field exhibits the best DOA estimation performance along with reduction in the computational cost.**

***Index Terms*—CNN, source localization, DOA, multi-scale aggregation**

I. Introduction

Many applications such as hands-free communication, teleconferencing, robot audition and distant speech recognition require information on the direction of sound sources in the acoustic environment. The relative direction of a sound source with respect to a microphone array is generally given in terms of the direction of arrival (DOA) of the sound wave originating from the source position. In most practical scenarios, this information is not available and the DOA of the sound source need to be estimated.

Compared to signal processing based approaches to this task, supervised learning approaches have the advantage that they can be adapted to different acoustic conditions via training. With the recent success of deep neural network based supervised learning methods for different signal processing related tasks, they have also become an attractive solution for DOA estimation [1]–[10].

Existing approaches mainly vary in terms of the input features that are utilized for the task of DOA estimation. Most of the earlier methods [1]–[5] involved a feature extraction step, where features similar to those used in classical signal processing based approaches, were given as an input to a deep neural network to learn the mapping from the features to the DOA of the sound sources.

The current authors proposed a convolutional neural network (CNN) based supervised learning method for broadband DOA estimation where the phase component of short-time Fourier transform (STFT) coefficients of the input signal were directly provided as input to the neural network [6]. Following this, other methods were also proposed that use the time-frequency representation of the recorded microphone signals as input to the neural network and utilized convolution layers in the early part of the proposed architecture to learn the required features from the input for the DOA estimation task [7], [8], [10].

In [10] a CNN based DOA estimation framework for the task of estimating the DOAs of multiple simultaneously active sound sources was proposed by the current authors. In that work, an experimental analysis of the number of required convolution layers for the best performance of the proposed framework was presented and it was found that for a uniform linear array (ULA) with M microphones, $M-1$ convolution layers are required due to the choice of small filters that learn the phase relations from neighboring microphones. Such a requirement limits the applicability for microphone arrays with large number of microphones, since it leads to a high computational cost due to the large number of convolution layers required to achieve a good DOA estimation performance. Also, a large number of convolution layers leads to a deep network which can lead to the vanishing gradient problem [11]. In [10], it was also shown that by simply reducing the number of convolution layers, the proposed method suffered from considerable degradation in performance.

In this paper, we propose to expand the receptive field of the filters by systematic dilation of the filters in each of the convolution layers. This dilation relaxes the requirement of $M-1$ convolution layers in [10]. The idea of using systematic dilated convolutions for multi-scale aggregation of contextual information in the feature space was first introduced in [12] for dense prediction tasks in computer vision. In [12], it was demonstrated that dilated convolutions allow

978-9-0827-9703-9/19/$31.00

for the exponential expansion of the receptive field of the filters without loss of resolution. In audio-related tasks, dilated convolutions have been mainly used in generative models for audio synthesis [13]. In this work, we utilize systematic dilated convolutions for aggressive expansion of the receptive field of the convolution filters such that phase information from all the microphones can be aggregated in fewer than $M-1$ convolution layers. The main contribution of the work presented here lies in showing the use of dilated convolutions for complexity reduction in a CNN based DOA estimation framework. Different strategies for the incorporation of dilated convolutions within the convolution layers are investigated for a specific microphone array and the reduction in computational cost that can be achieved is presented. We also investigate the use of larger filters as a possible solution to this problem. Finally, the DOA estimation performance of the proposed modifications to the CNN presented in [10] is compared to the original architecture.

II. DOA estimation with CNNs

In this section, a brief overview of the previously proposed CNN based framework [7], [10] for estimating the DOAs of multiple simultaneously active speakers is presented.

DOA estimation in [7], [10] is performed for signal blocks that consist of multiple time frames of the Short-time Fourier Transform (STFT) representation of the observed signals, where the block length can be chosen based on the application scenario.

The problem is formulated as a multi-label multi-class classification problem, where the complete DOA range is discretized to form the DOA class labels. The CNN learns to assign multiple DOA class labels to the input corresponding to each STFT time frame using a large amount of training data. In the test phase, the first task is to estimate the posterior probability of each DOA class given the input feature representation corresponding to a single STFT time frame. The assignment of each DOA class label is treated as a separate binary classification problem, assuming an independent source location model. The frame-level probabilities are subsequently averaged over a predefined number of time frames, hereafter referred to as a block. Finally, considering L sources, the DOA estimates are given by selecting the L DOA classes with the highest probabilities.

The input feature representation in the framework is the *phase map*, that was first introduced in [6]. The phase map, for the n-th time frame is formed by arranging the phase of the STFT coefficients for each time-frequency bin (n,k) and each microphone m into a matrix of size $K \times M$, where $K = N_f/2+1$ is the total number of frequency bins, upto the Nyquist frequency, at each time frame and M is the total number of microphones in the array.

Given the phase map of a single STFT time frame as the input, the CNN generates the posterior probability for each of the DOA classes. In the convolution layers of the CNN, small filters of size 1×2 are applied to learn the phase relations between neighboring microphones at each frequency sub-band separately. These learned features for each sub-band are then aggregated by two fully connected layers leading to the output for the classification task.

An important design aspect of the previously proposed architecture is the number of convolution layers. In [10], it was experimentally demonstrated that $M-1$ convolution layers are required to obtain the best DOA estimation performance for a microphone array with M microphones. This result can be attributed to the fact that by using small filters of size 1×2, with each subsequent convolution layer after the first one, for each sub-band, the phase correlation information from different microphone pairs are aggregated by the growing receptive field of the filters, and to learn from the correlation between all microphone pairs, $M-1$ convolution layers are required to incorporate this information into the learned features.

One of the main drawbacks of this requirement is that for arrays with large number of microphones, the number of required convolution layers becomes high leading to a large computational requirement, which can become practically infeasible. Therefore, in this work we investigate possible modifications to this previously proposed architecture to reduce the computational requirement without significant degradation in the DOA estimation performance.

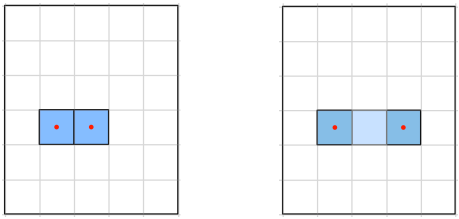
III. Receptive field expansion with Dilated Convolutions

The main reason for requiring $M-1$ convolution layers in the architecture proposed in [10] is the gradual aggregation of information in the feature space by the slowly growing receptive field of the small filters used in the framework. A possible solution for this problem is to use larger filters, however this can lead to an increase in the computational cost as well as the number of trainable parameters. In this work, we propose to incorporate systematic dilation of the convolution filters/kernels [12] to expand the receptive field of the filters with each convolution layer.

A. Dilated Convolutions

In dilated convolutions, based on the dilation factor, R, the applied filters are able to learn from distant elements in the input space while having the same number of filter elements as a generic discrete convolution filter that learns only from neighboring elements in the input space. In the case of 2D convolutions, different dilation factors can be applied along the different dimensions of the filter.

An illustrative example, in context of the filters used in our proposed architecture in [10], of the difference between the basic discrete convolution, also known as contiguous convolutions, and dilated convolution is shown in Fig. 1 where a dilation factor of 2 is applied along the column dimension of the filter. In Fig. 1(b), a dilated convolution operation is shown with the same number of filter elements as used in the contiguous convolution operation, shown in Fig. 1(a). The dilation factor mainly determines the size of the gap between the filter elements along a certain dimension. It should be noted that a dilation factor of 1 leads to the basic discrete convolution operation.



(a) Contiguous convolution

(b) Dilated convolution with dilation factor of $R = (1, 2)$

Figure 1: Illustrative example of contiguous and dilated convolution.

One of the main advantages of dilated convolutions is that with systematic inclusion of dilations, an exponential growth in the receptive field of the filters with each convolution layer can be achieved. For further details regarding dilated convolutions, we refer the readers to [12].

B. Design Considerations

The main aim in this work is to utilize the aggressive expansion of receptive field afforded by using dilated convolutions to reduce the requirement of $M - 1$ convolution layers within our previously proposed architecture. Keeping the number of elements in the filters constant, the main design choice is the dilation factor for each convolution layer.

One of the first things to note is that in the first convolution layer the dilation factor should be 1. This is required to avoid any loss of resolution in the feature space for the learned filters, since a dilation factor of greater than 1 restricts the filters from learning from phase relations between the neighboring microphones. In preliminary experiments, significant reduction in performance was observed when using a dilation factor of greater than 1 in the first layer.

The choice of dilation factor for the subsequent layers depends on the desired reduction in the number of convolution layers. In the convolution layers, the number of parameters is generally low since the size of the filters is significantly lower than the input dimensions and it relies on weight-sharing to learn similar patterns at different locations in the input feature space. However, the number of computations is generally high due to the striding of the filters through the complete input space. Therefore, with the reduction in the number of convolution layers, a higher decrease in the computational requirement can be observed compared to the reduction in the number of trainable parameters for the complete architecture. In Section IV-C, the reduction in computational cost as well as in the number of trainable parameters is discussed in the context of the proposed architecture in [10].

It should be noted that for the proposed architecture in [10], with a specific choice of number of convolution layers, the dilation factors of all the convolution layers should sum up to $M - 1$ such that the microphone dimension of the output feature map after the last convolution layers is 1, as once the microphone dimension is squashed to 1, no further 2D convolutions are possible. This keeps the number of trainable parameters manageable while also having the filters cover the whole microphone dimension of the input feature space.

IV. Experimental Validation

The design modifications proposed in the previous section are experimentally validated in this section with a specific microphone array for different simulated acoustic conditions. In [10], the CNN based method was shown to outperform two different traditional DOA estimation methods and was evaluated with real recordings as well. To have accurate ground-truth information, simulated data is used in the following experiments and due to space constraints, only the architecture proposed in [10] is used for comparison.

A. Experimental Setup

For the experimental investigation, we consider a uniform linear array (ULA) with $M = 8$ microphones with inter-microphone distance of 2 cm, and the input signals are transformed to the STFT domain using a DFT length of $N_f = 512$, with 50% overlap, such that $K = 257$. The sampling frequency of the signals is $F_s = 16$ kHz. To form the classes, we discretize the whole DOA range of a ULA, $[0°, 180°]$,with a $5°$ resolution to get $I = 37$ DOA classes, for both training and testing. All the presented objective evaluations are for the two speakers scenario.

The speech signals used for evaluation are taken from the LibriSpeech corpus [14]. During test, for a specific source-array setup in a room, a two speaker mixture is considered for all possible angular combination. This was done to avoid the influence of signal variation on the difference in performance for different acoustic conditions. Since the speech utterances can have different lengths of silence at the beginning, the central 0.8 s segment of the mixtures was selected for evaluation. Considering an STFT window length of 32 ms with 50% overlap, this resulted in a signal block of $N = 50$ time frames over which the frame-level probabilities are averaged for the final DOA estimation.

For evaluation, two different objective measures were used: Mean Absolute Error (MAE) and localization Accuracy (Acc.) [10]. As described in [10], an estimate is considered accurate if it lies within $5°$ of the true DOA. To be able to compute the accuracy with respect to both sources, it was therefore ensured that the angular distance between the two simultaneously active speakers is at least $10°$.

B. Compared Architectures

To experimentally investigate the difference in performance due to different design choices regarding the convolution filters, we compare the performances of the following architectures:

- CNN with contiguous convolutions proposed in [10], with different number of convolution layers ranging from 2 to $M - 1 = 7$. This is referred as the baseline architecture.
- CNN with contiguous convolutions with larger filters after the first convolution layer. We reduce the number of convolution layers to be 4, with filters of size 2 for the

Table I: Number of convolution layers, number of FLOPs and number of trainable parameters for the different compared architectures.

Architectures	Conv. Layers	FLOPs ($\times 10^6$)	FLOPs (w.r.t [10])	Tr. Parameters ($\times 10^6$)
[10]	7	53.14	1	8.75
F2342	4	35.24	0.66	8.84
D1123	4	32.08	0.60	8.73
D133	3	19.45	0.36	8.72

Table II: Configuration for generating test data for the experiments. All rooms are 3 m high.

Signal	Speech signals from LIBRI
Room size	Room 1: (4×7) m , Room 2: (9×7) m
Array positions in room	3 arbitrary positions in each room
Source-array distance	1.3 m for Room 1, 2.1 m for Room 2
RT_{60}	Room 1: 0.38 s , Room 2: 0.52 s
SNR	20 dB and 30 dB

first layer, followed by filters of size 3,4 and 2. This architecture is referred as F2342.

- CNN with dilated convolutions and 4 convolution layers. The dilation factors starting from the first convolution layer are 1,1,2 and 3. This architecture is referred as D1123.
- CNN with dilated convolutions and 3 convolution layers. The dilation factors starting from the first convolution layer are 1,3 and 3. This architecture is referred as D133.

In total, 9 different CNNs were trained for this experiment. The number of fully connected layers, as well as the activation functions were same for all the networks. All the networks were trained with the same amount of data. Multi-condition training with simulated training data for diverse acoustic conditions, as described in [10], was performed for robust performance in different acoustic scenarios.

C. Computation and Memory Complexity

In this section, we look at the computation and memory requirements for the different architectures introduced in the previous section. The computational requirement is presented in terms of the number of floating point operations (FLOPs) required for a single forward pass of the input phase map through the network. For each convolution layer, the FLOPs are computed as the product of three dimensions (height, width , depth) of the input map and the dimensions of the filters. For the fully connected layers, the number of FLOPS is given as a product of the input and the output vector lengths. The memory requirement is given in terms of the total number of trainable parameters (including bias terms) in each architecture.

The number of convolution layers, and the computation and memory requirements for the different compared architectures is shown in Table I. For the baseline architecture, the computation and memory requirement for only the network with $M - 1 = 7$ convolution layers is shown.

It can be seen in Table I, that the memory requirement for the different architectures are quite similar. In CNN architectures where the convolution layers are followed by fully connected layers, most of the trainable parameters are found in the layer connecting the output of the convolution layers to the fully connected layer. Since all the architectures have the same dimension of the feature maps at the output of the convolution layers and the same number of neurons in the adjoining fully connected layer, their memory requirements are similar. The network with larger filters (F2342) has the highest memory requirement due to the application of larger filters.

Though the memory requirement for the different architectures are similar, the number FLOPs for the networks with dilated convolutions (D1123, D133) and larger filters (F2342) are much lower than the previously proposed architecture (as shown in Table I). As explained in Section III-B, a decrease in the number of convolution layers leads to considerable decrease in the number of FLOPs. For the network D133, where higher dilation factor is applied in the early convolution layers for a more aggressive expansion of the receptive field of the filters, the number of FLOPs is almost a third of that of the baseline architecture. For the network with dilated convolutions with a more gradual expansion of the receptive field (D1123), the number of FLOPs is 60% of the baseline architecture. It can also be seen that by using larger filters, the number of FLOPs can be reduced by 35% for the microphone array setup considered here.

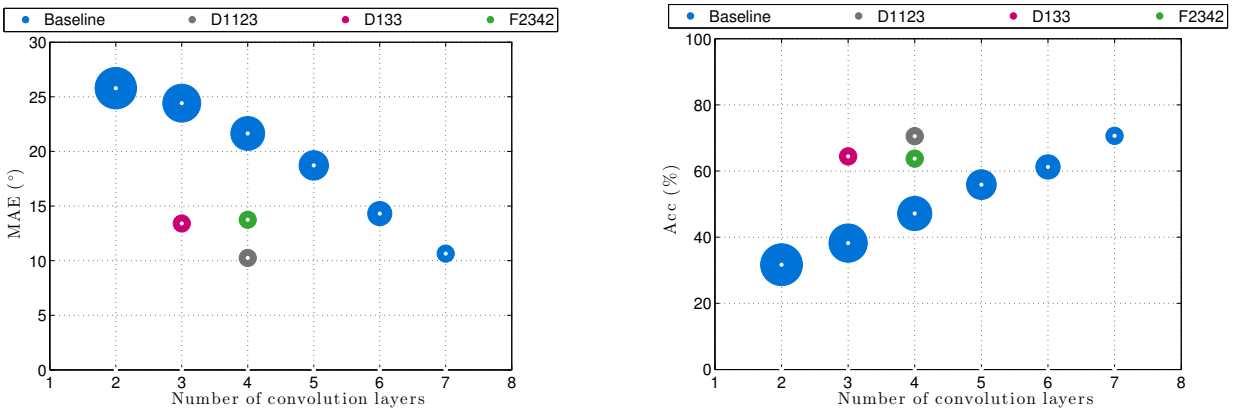
Therefore, by introducing the proposed modifications to the network, a considerable reduction in computational requirement can be achieved compared to the baseline architecture.

D. Results

The DOA estimation performance of the 9 different trained networks was evaluated with simulated data. The acoustic conditions used for test are shown in Table. II. Please note that there was no overlap between the acoustic configurations considered for training and testing.

The performance of the different compared architectures, in terms of both MAE and localization accuracy, is shown in Fig. 2. In the figures, the center of the circle markers correspond to the value of the objective measure and the area of the markers denote the number of trainable parameters for that specific architecture. Due to space constraints, we present results averaged over the different acoustic conditions given in Table II.

For the baseline architecture, it can be seen that by simply reducing the number of convolution layers, there is considerable degradation in performance (blue in Fig. 2). By using larger filters with 4 convolution layers (F2342, green in Fig. 2), an improvement of 8° in terms of MAE and 16% in terms of



(a) MAE (b) Accuracy

Figure 2: DOA estimation performance of the different compared architectures.

accuracy can be seen compared to the baseline architecture with 4 convolution layers. Using dilated convolutions and an aggressive receptive field expansion strategy (D133, pink in Fig. 2), slightly better performance than F2342 is achieved with only 3 convolution layers but compared to the baseline architecture with 7 layers, there is a loss of 7% in terms of accuracy and the MAE increases by 4°. The best performance is achieved by the D1123 network (gray in Fig. 2), which consists of 4 convolution layers and uses a gradual expansion of the receptive field.

From the results, we see that by using systematic dilations with a gradual expansion strategy (D1123), a performance similar to the baseline network with $M-1$ convolution layers is achieved for the considered scenario, with a reduction of 40% in FLOPs. With an aggressive expansion strategy for the receptive field of the filters (D133), the computation requirement can be reduced by 64%, however there is a 7% loss in accuracy and 4° increase in MAE compared to the baseline network with $M-1$ layers.

V. Conclusion

We proposed an approach to reduce the computational requirement of our previously proposed network for DOA estimation. In particular, we expanded the receptive field of the filters in the convolution layers by systematically dilating the filters. It was shown that by utilizing systematic dilation the computational cost can be reduced while keeping the memory requirement similar. Through experimental analysis with a specific microphone array, it was found that though an aggressive expansion of the receptive field of the filters leads to almost 65% reduction in the computational cost, it suffers from a 7% loss in accuracy compared to the original network. A more gradual expansion of the receptive field leads to a 40% reduction in computational cost and a negligible degradation in performance.

References

[1] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec 2017.

[2] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[3] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 405–409.

[4] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2814–2818.

[5] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 603–609.

[6] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2017.

[7] S. Chakrabarty and E. A. P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," in *ML4Audio Worskhop at Proc. Neural Information Processing Conf*, 2017.

[8] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2018.

[9] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tokyo, Japan, Sept. 2018.

[10] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[13] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.