RAW WAVEFORM BASED END-TO-END DEEP CONVOLUTIONAL NETWORK FOR SPATIAL LOCALIZATION OF MULTIPLE ACOUSTIC SOURCES

Harshavardhan Sundar[1], *Weiran Wang*[2*], *Ming Sun*[1], *Chao Wang*[1]

[1]Amazon.com Inc. [2]Salesforce Research

Email: {sundarhs,mingsun,wngcha}@amazon.com weiran.wang@salesforce.com

ABSTRACT

In this paper, we present an end-to-end deep convolutional neural network operating on multi-channel raw audio data to localize multiple simultaneously active acoustic sources in space. Previously reported deep learning based approaches work well in localizing a single source directly from multi-channel raw-audio, but are not easily extendable to localize multiple sources due to the well known permutation problem. We propose a novel encoding scheme to represent the spatial coordinates of multiple sources, which facilitates 2D localization of multiple sources in an end-to-end fashion, avoiding the permutation problem and achieving arbitrary spatial resolution. Experiments on a simulated data set and real recordings from the AV16.3 Corpus demonstrate that the proposed method generalizes well to unseen test conditions, and outperforms a recent time difference of arrival (TDOA) based multiple source localization approach reported in the literature.

Index Terms— Acoustic Source Localization, Deep Learning, Convolutional Neural Networks, Raw Waveform

1. INTRODUCTION

Acoustic source localization (ASL) pertains to the problem of localizing an acoustic source in space using only the audio data captured by an array of microphones. Historically, this problem of localizing a single acoustic source (single source localization or SSL in short) has attracted a number of signal processing based solutions [1–5]. Since the utility of an SSL algorithm is limited in realistic settings where multiple acoustic sources could be simultaneously active, several authors have proposed signal processing based algorithms for multiple source localization (MSL) [6–13]. A major advantage of treating ASL as a signal processing problem is that, there is no training phase and thus no training data is required. The disadvantages, however, stem from the fact that, these algorithms can be more complicated to deploy than a deep neural network, for which standard deep learning libraries with a wide range of supported hardware are available [14]. Furthermore, making these signal processing based algorithms robust to specific distortions requires algorithm-specific insights.

ASL can also be treated as a machine learning problem wherein the goal is to map the location bearing features derived from the microphone signals to the coordinates of the source. More commonly, this problem is cast as a classification problem by discretizing the enclosure into a grid of possible source locations [15–18]. In [15], the authors employ a neural network trained to map the directional features derived from the short-time Fourier transform (STFT) of the multi-channel audio to a discrete source location. In [16] the authors use the magnitude and phase STFT to localize multiple sources using deep convolutional-recurrent neural networks (CRNN). In [17] the multi-channel STFT phase is used as input to a deep convolutional neural network (CNN). In [18]the authors propose to employ deep feed-forward/convolutional neural networks to predict likelihood-based encoding of angular location of sources, taking generalized cross correlation (GCC) with phase transform (PHAT) and GCC computed on sub-bands of a gammatone filter bank as inputs. In all of the above approaches, the resolution of the algorithm to localize a single source is limited to the resolution of the grid which is typically between $5^\circ - 10^\circ$.

More recently, [19] have explored the possibility of an end-to-end system mapping the raw multi-channel audio from a microphone array directly to the source coordinates. A major drawback of this approach is that there is no straight forward way to extend it to localize multiple sources. Increasing the number of output nodes to estimate the coordinates of each source leads to the well known permutation problem, which also manifests in time difference of arrival (TDOA) based MSL approaches [6, 20]. More importantly, in practice, the number of active sources changes over time, and it is not clear how to extend such an approach to realistic scenarios.

We present a deep convolutional neural network (CNN) to map the raw audio data from a microphone array to the 2D coordinates of all sources, without any pre-/post-processing of inputs/outputs. To the best of our knowledge, this is the first end-to-end deep learning system for MSL. Based on the insights from [6] for avoiding the permutation problem, we propose a joint Coarse-Fine localization strategy, in which a source is associated with a coarse and a fine location. Instead of outputting the source coordinates directly, we propose to output encoded coordinates based on the coarse location of the source. Such a design facilitates simultaneous localization of multiple sources with arbitrary resolution. Additionally, it also ensures that a source within a specific coarse location is always associated with a specific output node, thereby avoiding the permutation problem. We train a "SampleCNN" architecture based on residual connections [21] with the proposed output encoding layer to detect the coarse regions containing an active source, while simultaneously estimating the source location finely within each such active coarse region. This is achieved using a classification cost to measure the discrepancy between actual and detected coarse regions, and a regression cost to finely localize sources within each coarse region. The network is trained to minimize a combination of these two costs. Unlike the approaches in [16, 17], our method does not simply treat the MSL problem as a multi-label classification task, but as a joint multi-label classification and regression task. Initial results on a simulated data set and on real recordings from the AV16.3 corpus show that the proposed end-to-end approach significantly outperforms a recent signal processing solution based on GCC/TDOA estimation [6], especially when there are two or more active sources.

Notations Scalar variables are denoted in lowercase, constants in uppercase, vectors in boldface-lowercase and matrices in boldface-uppercase. We use a uniform circular array (UCA) of microphones. The number of microphones in the array is denoted by M and the number of active sources is denoted by K. The locations of the microphones are assumed to be known and are denoted by the position vectors $\mathbf{p}_i = [x_i, y_i, z_i]\,; 1 \leq i \leq M$. The M microphone signals are represented as $w_i[n]; 1 \leq i \leq M$. F_s denotes the sampling frequency of the digitized microphone signals.

*Work done while Weiran Wang was at Amazon.com Inc.

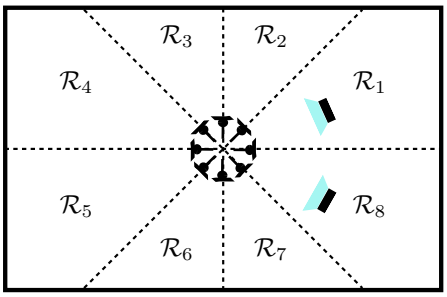

Fig. 1: *A 2D Schematic of the enclosure with $L = 8$ sector-like partitions and a UCA with $M = 8$ microphones. Two acoustic sources are present in regions $\mathcal{R}_1$ and $\mathcal{R}_8$.*

2. DEEP CNN FOR LOCALIZATION OF MULTIPLE ACOUSTIC SOURCES

2.1. Encoding Spatial Coordinates of Multiple Sources

A network designed to output the coordinates of multiple sources presents a permutation problem even when localizing stationary sources over multiple segments of data. This permutation problem arises when it cannot be ensured that the coordinates of a particular source always appears on the same output node of the network. To circumvent this problem, we utilize the insights reported in [6], and partition the enclosure into L sector-like coarse regions denoted by $\{\mathcal{R}_\ell\}_{\ell=1}^{\ell=L}$ centered around the UCA. A region is said to be active if it contains at least one source and is said to be inactive otherwise. We assume that there can be at most one active source in any active region. Consequently, the proposed network can localize up to L simultaneously active sources. Note that this assumption in general is not restrictive as the regions can be made smaller, justifying the above assumption of one source per region. The 2D schematic of a representative enclosure partitioned into $L = 8$ sector-like regions, along with a UCA of $M = 8$ microphones and two sources - one in $\mathcal{R}_1$ and the other in $\mathcal{R}_8$ is shown in Figure 1.

Let the 2D coordinates of a source in $\mathcal{R}_\ell$, located at $\mathbf{p}_{S_\ell}$, be specified in terms of the radial distance (d_ℓ) and azimuthal angle (θ_ℓ) computed with respect to the center of the microphone array $(\mathbf{p}_0)$. Then,

$$d_\ell = \|\mathbf{p}_{S_\ell} - \mathbf{p}_0\|, \quad \text{and} \quad \theta_\ell = \angle(\mathbf{p}_{S_\ell} - \mathbf{p}_0), \tag{1}$$

where $\angle$ denotes the azimuthal angle operator and $\|.\|$ denotes the Euclidean distance between the two vectors. Instead of representing the source coordinates directly, we encode the source coordinates in each sector as:

$$\tilde{d}_\ell = \frac{d_\ell - d_\ell^{\min}}{d_\ell^{\max} - d_\ell^{\min}}, \quad \text{and} \quad \tilde{\theta}_\ell = \frac{\theta_\ell - \theta_\ell^{\min}}{\theta_\ell^{\max} - \theta_\ell^{\min}}, \tag{2}$$

where the limits $\left(d_\ell^{\min}, d_\ell^{\max}\right)$ represent the minimum and maximum possible radial distance, and $\left(\theta_\ell^{\min}, \theta_\ell^{\max}\right)$ represent the minimum and maximum possible azimuthal angle for points in $\mathcal{R}_\ell$. As a consequence of the parameterization (2), we have $0 \leq \tilde{d}_\ell, \tilde{\theta}_\ell \leq 1$, as targets for training neural networks.

For a given input sample, the network is designed to output 3 quantities corresponding to each coarse region $\mathcal{R}_\ell; 1 \leq \ell \leq L$:

- $\hat{r}_\ell \triangleq \Pr(\mathcal{R}_\ell \text{ is active})$
- $\left(\hat{\tilde{d}}_\ell, \hat{\tilde{\theta}}_\ell\right)$ - the encoded 2D coordinates of a source in $\mathcal{R}_\ell$,

resulting in a total of $3 \times L$ outputs. With this design, the network is capable of simultaneously localizing up to L sources while avoiding the permutation problem altogether.

2.2. Overall Network Architecture

Based on the recent success in using raw-audio based convolutional networks as reported in [21], we propose a similar network architecture using residual and "squeeze-and-excitation" blocks as shown in Figure 2. It is well-known that the cross-correlation of microphone signals contain location information [6, 22, 23]. Thus, we use only 1D convolutional filters which shall be able to extract the required location information.

The multi-channel audio of duration T seconds forms the input to the network. Consider one such multi-channel audio sample $W \in \mathbb{R}^{T.F_s \times M}$. W is first processed by a "wave-norm" layer wherein, each input waveform is normalized to have a maximum amplitude of unity. This was found to be useful in our initial experiments. The normalized raw-audio samples are then fed to the basic convolutional block shown in Figure2(a), containing a single convolutional layer with 64 1D convolutional filters, followed by the batch normalization layer and rectified linear unit (ReLU) non-linearity as prescribed in [21, 24]. The output from the basic block is down-sampled by a factor of 3, and fed to a series of residual squeeze and excitation blocks containing 2 convolutional layers each and abbreviated as "ReSE-2" blocks. The operations performed by each ReSE-2 block is summarized in Figure 2(b) and 2(c). In each ReSE-2 block the input passes through 2 convolutional layers with 128 1D convolutional filters and is down-sampled by a factor of 3. The 2D output of the final ReSE-2 block is then transformed into a 256 dimensional vector, through global max pooling layer and 2 fully connected layers, which is used for generating the final classification outputs (referred to as coarse location predictions), and regression outputs (referred to as fine location predictions), to detect active regions and localize the sources finely within each active region, respectively.

For a threshold based detection, the active regions are first identified as the ones with coarse location predictions higher than a threshold. Let $\mathcal{D} \triangleq \{\ell : \hat{r}_\ell >= \epsilon\}$ be the set of indices corresponding to active regions, where ϵ is the detection threshold. For each detected region, the source location $\left(\hat{d}_{\ell^*}, \hat{\theta}_{\ell^*}\right)$ is computed from the encoded fine location predictions - $\left(\hat{\tilde{d}}_{\ell^*}, \hat{\tilde{\theta}}_{\ell^*}\right)$ $\forall \ell^* \in \mathcal{D}$ as:

$$\hat{d}_{\ell^*} = \hat{\tilde{d}}_{\ell^*}.\left(d_\ell^{\max} - d_\ell^{\min}\right) + d_\ell^{\min} \tag{3}$$

$$\hat{\theta}_{\ell^*} = \hat{\tilde{\theta}}_{\ell^*}.\left(\theta_\ell^{\max} - \theta_\ell^{\min}\right) + \theta_\ell^{\min} \tag{4}$$

If the number of active sources is K, then $\mathcal{D}$ constitutes K active region indices corresponding to the top K scores in $\{\hat{r}_\ell\}_{\ell=1}^{L}$. The source in each active region is then estimated using (3) and (4).

The proposed approach is referred to as Sample based Multiple Encoded Source Location Predictor (SMESLP).

2.3. Loss Function

The training data for the proposed network architecture is a triplet - $(\boldsymbol{W}, \boldsymbol{R}, \boldsymbol{D})$. $\boldsymbol{W} = \left[\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \cdots, \boldsymbol{w}^{(J)}\right]$ denotes the raw-audio samples input to the network where $\boldsymbol{w}^{(j)} \in \mathbb{R}^{T.Fs \times M}$ and J indicates the number of training samples. $\boldsymbol{R} = \left[\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \cdots, \boldsymbol{r}^{(J)}\right]$ represents the ground truth of the coarse region labels where $\boldsymbol{r}^{(j)} = \left[r_1^{(j)}, r_2^{(j)}, \cdots, r_L^{(j)}\right]^T$, $r_\ell^{(j)} = 1$ if $\mathcal{R}_\ell$ is an active region in the j^{th} training sample and 0 otherwise. $\boldsymbol{D} = \left[\boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, \cdots, \boldsymbol{d}^{(J)}\right]$ represents the encoded 2D location of the source in each of the L regions, where $\boldsymbol{d}^{(j)} = \left[\left(\tilde{d}_1^{(j)}, \tilde{\theta}_1^{(j)}\right), \left(\tilde{d}_2^{(j)}, \tilde{\theta}_2^{(j)}\right), \cdots, \left(\tilde{d}_L^{(j)}, \tilde{\theta}_L^{(j)}\right)\right]^T$. Encoded coordinates - $\left(\tilde{d}_\ell^{(j)}, \tilde{\theta}_\ell^{(j)}\right)$ are related to the corresponding absolute coordinates - $\left(d_\ell^{(j)}, \theta_\ell^{(j)}\right)$ $\forall 1 \leq \ell \leq L$ as in (2).

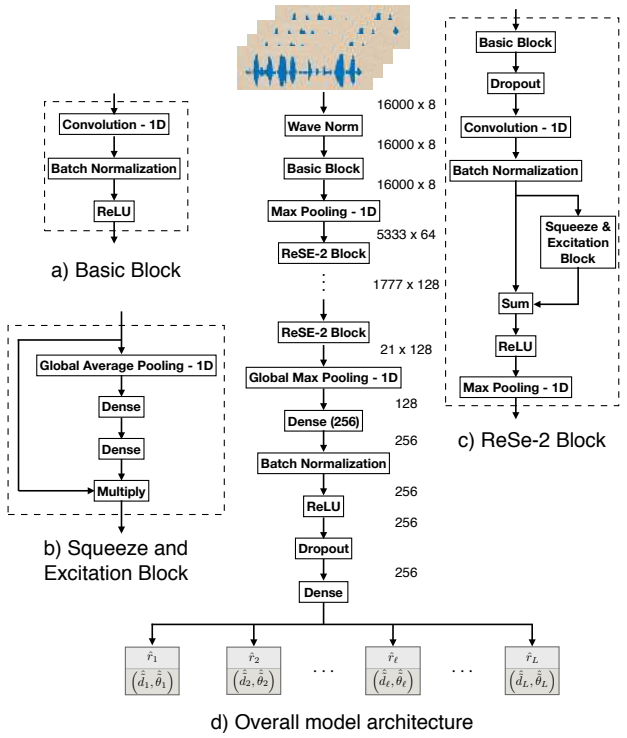

Fig. 2: *A block diagram of the model architecture. The tensor sizes at the output of each layer is also indicated.*

For the j^{th} sample, the coarse location prediction from the network is represented as $\hat{\boldsymbol{r}}^{(j)}$ and the fine location prediction is represented as $\hat{\boldsymbol{d}}^{(j)}$. We define the training loss for the j^{th} sample - $\mathcal{L}^{(j)}$ as a weighted sum of the coarse localization loss - $\mathcal{L}^{(j)}_{\text{Coarse}}$ and the fine localization loss - $\mathcal{L}^{(j)}_{\text{Fine}}$, i.e.,

$$\mathcal{L}^{(j)} = \alpha \cdot \mathcal{L}^{(j)}_{\text{Coarse}} + \beta \cdot \mathcal{L}^{(j)}_{\text{Fine}},$$

where $\mathcal{L}^{(j)}_{\text{Coarse}}$ is the multi-label classification loss between the predicted regions and the actual regions averaged over the regions:

$$\mathcal{L}^{(j)}_{\text{Coarse}} = -\frac{1}{L}\sum_{\ell=1}^{L}\left[r^{(j)}_{\ell}\log\left(\hat{r}^{(j)}_{\ell}\right) + \left(1-r^{(j)}_{\ell}\right)\log\left(1-\hat{r}^{(j)}_{\ell}\right)\right],$$

and $\mathcal{L}^{(j)}_{\text{Fine}}$ is the average Euclidean distance between the actual and predicted encoded coordinates of sources in active regions:

$$\mathcal{L}^{(j)}_{\text{Fine}} = \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{1}_{\left\{r^{(j)}_{\ell}=1\right\}}\sqrt{\left(\tilde{d}^{(j)}_{\ell}-\hat{\tilde{d}}^{(j)}_{\ell}\right)^2+\left(\tilde{\theta}^{(j)}_{\ell}-\hat{\tilde{\theta}}^{(j)}_{\ell}\right)^2},$$

where $\mathbb{1}_{\left\{r^{(j)}_{\ell}=1\right\}}$ is the indicator function for $\mathcal{R}_{\ell}$ being active. The overall training loss is computed as $\mathcal{L} = \frac{1}{J}\sum_{j=1}^{J}\mathcal{L}^{(j)}$. The SMESLP networks are trained using Tensorflow[1] and Keras[2].

3. EXPERIMENTS

Although there are a few publicly available data sets like the LOCATA challenge data set [25], because of our choice of the microphone array and the requirement of a large data set to train the proposed network, we resort to simulated data for training our model. However, in section 3.3 we test this network on real recordings.

3.1. Simulated Data set Details and Performance Metrics

The microphone signals are simulated by suitably transforming clean speech data from dialect 8 (DR8) of the TIMIT database [26], sampled at 16 kHz. Both anechoic and reverberant data (with a 60 dB reverberation time RT60 of 300 ms) are simulated for training and evaluation of the proposed network. The anechoic data set is created by superposition of appropriately shifted and attenuated clean speech signals for each active source based on the source-microphone distance. The reverberant data are created by convolving clean speech signals with room impulse response (RIR) generated using [27] based on the Image method [28]. For RIR generation, an enclosure of size 6m $\times$7.5m $\times$4.5m is used. The enclosure is divided into $L = 8$ sector-like regions with equal azimuthal angular width of 45°. A UCA with $M = 8$ microphones and a radius of 10 cm is used and is placed at the center of the enclosure. In each region $\mathcal{R}_{\ell}; 1 \leq \ell \leq 8$, 312 points are randomly sampled [29, 30] out of which 250 are chosen for training, and 36 each for validation and testing. Up to 3 simultaneous sources are simulated. Each training/evaluation sample is chosen to be of 1s duration. Care is taken to ensure similar number of samples are generated for different number of simultaneously active sources. The data set statistics are summarized in Table 1. We analyze the performance of the localization systems using two types of measures. One measure to capture how well the proposed approach is able to coarsely localize the sources to be within the regions - $\mathcal{R}_{\ell}; 1 \leq \ell \leq 8$. This is done using Hamming accuracy (ACC), also known as the Jaccard index, which is a suitable metric for analyzing multi-label classification systems [31]. It is defined as ACC $= \frac{\|T \bigcap P\|}{\|T \bigcup P\|}$ with T and P representing the set of true and predicted labels respectively. The other measure is to capture how well the proposed approach is able to finely localize the detected sources, which is done using the absolute direction of arrival (DOA) error between the actual and estimated source locations. In order to conform with the existing literature, the performance on the AV16.3 corpus is measured in terms of the root-mean-squared-error (RMSE) of direction-of-arrival (DOA) and the percentage of non-anomalous frames defined as the percentage of $1s$ frames in which the algorithms estimate the source within a DOA error of 10°.

Table 1: Simulated Data set statistics.

Acoustic Condition	Train	Validation	Test
Anechoic	33,356	443	414
Reverb	34,196	460	456

3.2. Performance on Simulated Data

We compare performance of the proposed networks trained on anechoic data only called SMESLP-Anechoic, and trained on reverberant data only called SMESLP-Reverb, with a baseline. The baseline is the GCC/TDOA based approach "I-IDIR-UCA" by [6]. Since I-IDIR-UCA is also a region based approach, we compare its coarse and fine localization performances with the proposed networks in one, two and three active sources scenarios. Figure 3 (left panel) shows the Hamming score (ACC) for the three techniques on anechoic data (top row) and on reverberant data (bottom row). In a single source scenario, all three techniques perform coarse localization perfectly. On the anechoic test set, the I-IDIR-UCA has a slight degradation in performance with increase in number of sources, while this degradation is more pronounced on the reverberant test set. SMESLP-Anechoic and SMESLP-Reverb have much lesser degradation on the number of sources. Interestingly, SMESLP-Anechoic generalizes well to the reverberant test set, while SMESLP-Reverb generalizes well to the anechoic test set. Closer examination of the results indicates that as reported in [6], I-IDIR-UCA does well on sources lying closer to the center of each region and is less robust to sources lying closer to the region boundaries. Our SMESLP approach is more robust to source lying closer to the boundary due to dense sampling of points for training.

Figure 3 (right panel) shows the absolute DOA error for the three approaches on anechoic (top row) and reverberant (bottom

[1] https://www.tensorflow.org/
[2] https://keras.io/

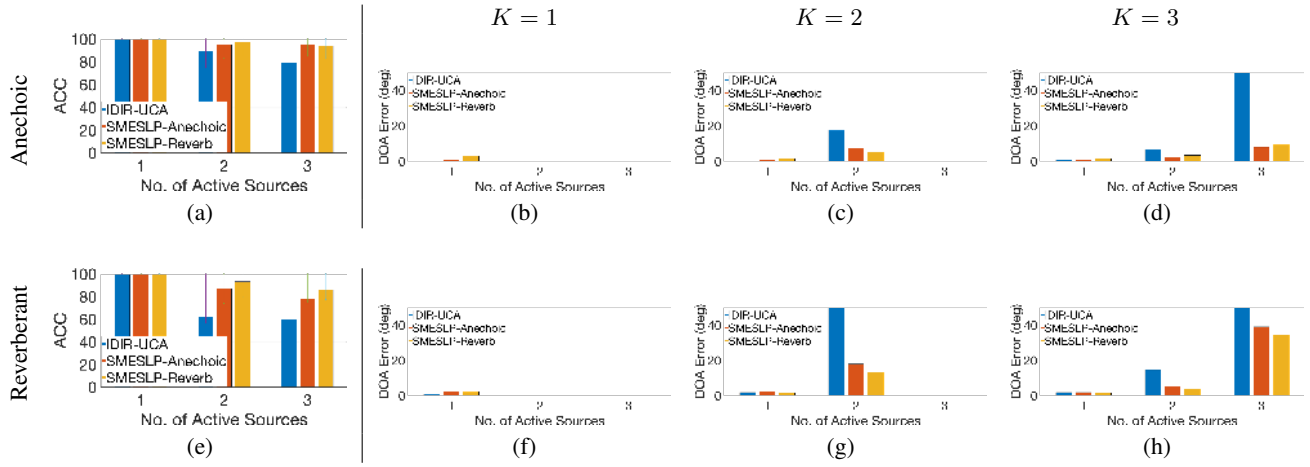

Fig. 3: *(Color Online): Left panel: Coarse localization performance measured using Hamming score (ACC) for the 3 techniques. Right panel: Absolute DOA error in* $\deg(^\circ)$ *for* $K = 1, 2$ *and* 3 *active sources. The top row corresponds to the performance evaluated on the anechoic test set and the bottom row corresponds to the performance on the reverberant test set.*

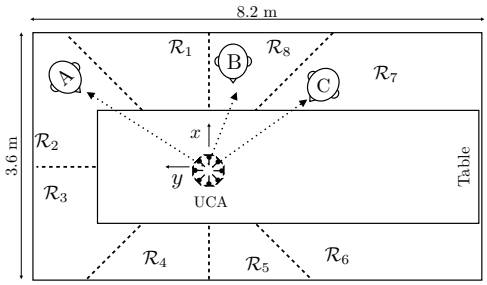

Fig. 4: *(Figure taken from [6])2D schematic of the localization setup used in the real recordings from the sequence* `seq37-3p-0001` *of the AV16.3 corpus [32]. Following [6], the enclosure is partitioned into 8 regions* $\mathcal{R}_1 - \mathcal{R}_8$*, the boundaries of which are shown in dashed lines.*

row) test sets with different number of active sources. Overall the proposed SMESLP approaches clearly outperform I-IDIR-UCA in the two and three active source scenario, with much smaller DOA error for localizing all the active sources.

3.3. Performance on AV16.3 Corpus

In this section, we analyze the performance of the proposed network on real recording sequence `seq37-3p-0001` from the AV16.3 corpus [32] recorded in the IDIAP smart room [33]. In this recording, three speakers A, B, and C converse from an azimuthal angular location of 74°, 353.5°, and 309.6°, respectively. While speakers B and C are stationary, speaker A moves to an angle of 62.16° and then to 314.1° over the duration of the sequence.

Similar to [6], we partition the enclosure into $L = 8$ regions as shown in Figure 4. The proposed network is compared with two other algorithms I-IDIR-UCA [6] and circular harmonics beamforming (CHB) [34]. We fine-tune the SMESLP-Anechoic network with data from AV16.3 corpus for 50 epochs with a lower learning rate of 0.0001. Of the 510 samples in the sequence `seq37-3p-0001`, 110 are used for fine tuning, 10 for validation and the remaining 380 for testing. In order to prevent the network from forgetting what it has learnt on the simulated data sets, we also include 100 random samples each from anechoic and reverberant data sets. Table 2 shows the mean absolute DOA error for the proposed approach and RMSE of DOA for comparing the proposed approach with the other two techniques. The percentage

Table 2: DOA Error and Percentage of non-anomalous frames (indicated within parentheses) in real recordings for the three approaches being compared.

	Sp. B	Sp. B, C	Sp. A, B, C
	Absolute DOA Error		
SMESLP	**1.13°(100%)**	**1.96°(97.95%)**	**2.05°(100%)**
	RMSE DOA Error		
SMESLP	**1.45°(100%)**	**2.33°(97.95%)**	**2.33°(100%)**
I-IDIR-UCA [6]	1.00° (92%)	1.83° (79 %)	4.1° (60%)
CHB [34]	1.18 °	2.00 °	2.98 °

of non-anomalous frames for the three approaches is also shown. Although the RMSE of DOA for SMESLP is slightly higher for one and two sources, in terms of the percentage of non-anomalous frames, the SMESLP approach outperforms I-IDIR-UCA [6]. For localizing three active speakers SMESLP approach outperforms the other two techniques.

4. CONCLUSIONS

An end-to-end deep CNN operating on multi-channel raw audio data is proposed to address the problem of localizing multiple acoustic sources in space. The main contribution of the paper is in designing an output layer to handle localizing multiple source while avoiding the permutation problem in localizing/tracking multiple sources over time. To the best of our knowledge, this is the first time an end-to-end approach is proposed for localizing multiple acoustic sources operating on raw multi-channel audio data. Evaluations on simulated data show that the proposed approach (SMESLP) trained on anechoic data generalizes to reverberant data and vice-versa. On real recordings from AV16.3 corpus, with fine-tuning on a small amount of data, the SMESLP approach clearly out-performed existing state-of-the-art approaches even with three active sources.

Deploying the SMESLP network in a completely different enclosure configuration from the one used for training, would require a small amount of fine-tuning data in order to achieve an acceptable level of performance. Overall, the SMESLP approach significantly reduces the domain knowledge required for deploying a multiple source localization system as compared to existing signal processing based approaches in the literature [6, 8, 9, 34] and can be deployed easily using existing deep learning frameworks.

5. REFERENCES

[1] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE Intl. Conf. Acoust. Speech and Signal Process.*, 2000, vol. 2, pp. 909–912.

[2] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown Univ., 2000.

[3] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 4, pp. 1327 – 1339, May 2007.

[4] K. W. K. Lui, F. K. W. Chan, and H. C. So, "Accurate time delay estimation based passive localization," *Signal Process.*, vol. 89, no. 9, pp. 1835–1838, Sep. 2009.

[5] R. De Mori, *Spoken Dialogues with Computers*, Academic Press, Inc., Orlando, FL, USA, 1998.

[6] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "TDOA-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 26, no. 11, pp. 1976–1990, Nov. 2018.

[7] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. Audio, Speech, and Music Process.*, 2010.

[8] G. Lathoud and J. Odobez, "Short-term spatio-temporal clustering applied to multiple moving speakers," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1696–1710, July 2007.

[9] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[10] H. Sawada, R. Mukai, S. Araki, and S. Malcino, "Multiple source localization using independent component analysis," in *Proc. IEEE Antennas and Propagation Soc. Intl. Symp.*, July 2005, vol. 4B, pp. 81–84.

[11] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *Proc. Intl. Conf. Acoust. Speech and Signal Process.*, April 2009, pp. 233–236.

[12] B. Loesch, S. Uhlich, and B. Yang, "Multidimensional localization of multiple sound sources using frequency domain ICA and an extended state coherence transform," in *Workshop on Stat. Signal Process.*, Aug. 2009, pp. 677–680.

[13] P. Teng, A. Lombard, and W. Kellermann, "Disambiguation in multidimensional tracking of multiple acoustic sources using a Gaussian likelihood criterion," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process.*, Mar. 2010, pp. 145–148.

[14] Giang Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Álvaro López García, I. Heredia, P. Malík, and L." Hluchý, "Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, Jun. 2019.

[15] 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. Acoust. Speech and Signal Process.*, Mar. 2016, pp. 405–409.

[16] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1462–1466.

[17] S. Chakrabarty and E. A. P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," *CoRR*, vol. abs/1712.04276, 2017.

[18] W.He, P. Motlícek, and J. M. Odobez, "Deep neural networks for multiple speaker detection and localization," *CoRR*, vol. abs/1711.11565, 2017.

[19] J. M. Vera-Diaz, D.Pizarro, and J. M. Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signal to source position coordinates," *CoRR*, vol. abs/1807.11094, 2018.

[20] W. K. Ma, B. N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, 2006.

[21] T. Kim, J. Lee, and J. Nam, "Comparison and analysis of samplecnn architectures for audio classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 285–297, May 2019.

[22] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[23] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[25] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, July 2018, pp. 410–414.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[27] E. A. P. Habets, "Room impulse response (RIR) generator," Sep. 2010.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer*, vol. 65, no. 4, pp. 943–950, 1979.

[29] D.P. Kroese, T. Taimre, and Z.I. Botev, "Handbook of monte carlo methods," pp. 240—-244, 2011.

[30] David E. Kaufman and Robert L. Smith, "Direction choice for accelerated convergence in hit-and-run sampling," *Operations Research*, vol. 46, no. 1, pp. 84—-95, 1998.

[31] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," *Advances in Knowledge Discovery and Data Mining*, pp. 22–30, 2004.

[32] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, S. Bengio and H. Bourlard, Eds., Berlin, Heidelberg, 2005, pp. 182–195, Springer.

[33] D. Moore, "The IDIAP smart meeting room," in *IDIAP Communication COM-02-07*, 2002.

[34] A. M. Torres, M. Cobos, B. Pueo, and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1511–1520, 2012.