MDPI

Article

Sound Source Localization Using Graph Regularized Neural Network

Firstname Lastname [1,†,‡], Firstname Lastname [1,‡] and Firstname Lastname [2,*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3
‡ These authors contributed equally to this work.

Version June 4, 2020 submitted to Journal Not Specified

Abstract: A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: Place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: Describe briefly the main methods or treatments applied; (3) Results: Summarize the article's main findings; and (4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.)

1. Introduction

Sound source localization is an increasingly important component in teleconferencing, autonomous driving, security and human-computer interaction systems.

It might be desired to use compact microphone arrays that have as few microphone elements as possible. By "compact" we here imply that the dimensions of the microphone array are much smaller than the dimensions of source localization space.

There are plenty of source direction of arrival estimation methods utilizing such compact arrays [], but only a few methods are offered for 2D or 3D source localization that can estimate both DoA and the distance or the Cartesian coordinates of the sound source.

Camera aiming can be inaccurate due to the parallax error if the camera and the microphone array are not rotating on the same axis. To accurately steer the camera to a certain point in space in such setup, a complete set of coordinates, either Cartesian or polar, of the target point is needed. Thus, knowing only the direction of arrival (DoA) of the sound source would not be sufficient. At least two DoAs are needed to describe the source position on a plane.

The Steered Response Power - Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. However, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue [1].

The performance of SRP-PHAT-based source localization algorithms deteriorate considerable when compact microphone arrays are used [].

Learning-based sound source localization methods might be further advantageous in such circumstances.

There are several learning-based source localization approaches, based on either semi-supervised [2] or supervised [3,4] learning paradigms. In both of these approaches to work, a set of acoustic features from known sound source positions (the labeled dataset) is needed.

Labeled feature acquisition is very costly. It is relatively easy to obtain a large dataset of unlabeled audio features. Considering our setting, it is relatively easy to collect a large amount of acoustic features without labels, and it is very tedious to provide labels (in our case – the coordinates of the sound source) for such data.

In this article we present a method to localize in two dimensions a sound source using a two compact microphone arrays.

2. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

> This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

2.1. Subsection

2.1.1. Subsubsection

Bulleted lists look like this:

- First bullet
- Second bullet
- Third bullet

Numbered lists can be added as follows:

1. First item
2. Second item
3. Third item

The text continues here.

2.2. Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

Figure 1. This is a figure, Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Text
Text

Table 1. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	Title 2	Title 3
entry 1	data	data
entry 2	data	data

Text
Text

2.3. Formatting of Mathematical Components

This is an example of an equation:

$$a + b = c \tag{1}$$

Please punctuate equations as regular text. Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:

Proof of Theorem 1. Text of the proof. Note that the phrase 'of Theorem 1' is optional if it is clear which theorem is being referred to. □

The text continues here.

3. Discussion

Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

4. Materials and Methods

In this section, we provide a theoretical background for our investigation.

ANN-based sound source localization can be performed using various acoustic features that depend on the sound source, either relative to the microphone array(s) or the acoustic enclosure.

It can be assumed, that acoustic features are spatially smooth, that is, features obtained at spatially close source positions are also close in feature space. The relative distance or affinity of the features can be simply estimated by calculating the Euclidean distance between feature vectors.

It can be assumed, that high dimensional acoustic features lie on a low-dimensional manifold, embedded in a high-dimensional feature space. Since the acoustic features are only dependent on the coordinates of the sound source, it is expected that the manifold would represent the spatial relations between the nearby acoustic features. We consider that that the affinity matrix of the low-dimensional embeddings of the acoustic manifold represent the *graph* of the acoustic features.

While the obtained embeddings of the acoustic manifold might represent the relative spatial relations between the acoustic features, it is not tied to physical properties and it also might be very non-linear. The translation of embedded space to physical space must be done in a separate step using a nonlinear regression method.

When ANN is utilized to obtain the sound source position via regression using acoustic features, it might be expected that the predictions of the ANN would also exhibit spatial smoothness. Feature manifold could be used during the training of the ANN to ensure that the ANN learns the relation between the acoustic features and the source positions while also retaining the spatial smoothness.

This spatial smoothness as well as the awareness of the relative spatial positions of the acoustic features is especially important when a semi-supervised learning strategy is involved. Using a training dataset that contains both labeled and unlabeled samples, knowledge of the relative distances between unlabeled and labeled samples might help to train the regressor to predict source locations for the unlabeled samples based on their manifold distance to the labeled features.

Our method is comprised of two stages.

1. The low-dimensional embedding of an acoustic feature manifold is obtained from a combined dataset of labeled and unlabeled samples. This manifold represents relative distances between acoustic feature in a low-dimensional embedded space.
2. A neural network is trained on the combined dataset, using a loss function that consist of a supervised loss (calculated only for labeled samples) and a graph loss (calculated for all samples, considering k nearest neighbors). Supervised loss ensures that the regressor is able to learn accurate relations between the acoustic features and the source positions. Graph loss ensures that the source position predictions remain spatially smooth.

Considering that the combined dataset consist of relatively low number of labeled samples and vast amounts of unlabeled samples, the supervised loss acts as a mean to "straighten" the manifold while the graph loss is used to infer the labels for the unlabeled samples based on their distance in feature (or embedded) space to the labeled samples.

4.1. Acoustic Features

We have considered several types of acoustic features, that are discussed further.

4.1.1. Time Difference of Arrival

Time Difference of Arrival (TDoA) is a trivial acoustic feature, that can be estimated using GCC-PHAT. Knowing the TDoA for several non-colinear (or non-parallel?) microphone pairs, it is possible to estimate the position of the sound source using triangulation (trilateration).

While this would be a simple and straightforward method, the accurate TDoA estimation becomes very tricky in reverberant or noisy environments. Moreover, the TDoA contains only very little information about the distance between the sound source and the microphone pair (just one value per pair). For a microphone array with 4 elements, that's only 6 values. TDoA does not explicitly contain any information about the structure of reflections withing the enclosure, nor the geometry or acoustic properties of the enclosure; it only depends on the relative source position with respect to the microphone array(s).

4.1.2. Room Impulse Response and Room Transfer Function

It is assumed that high-dimensional acoustic features, such as room impulse response (RIR) or room transfer function (RTF) contain a unique fingerprint of sound source and microphone positions within an enclosure. This is because the structure of room reflections is unique for every source position and every microphone position (theoretically, there might be some cases when same RIR is obtained for more than one combination of microphone and sound source positions, but this is probably possible in ideal room, which exhibit point symmetry around the center of the room; in real rooms this is impossible; also the microphones must be also placed symmetrically in the enclosure for this effect to occur).

While the RIRs and RTFs contain enough information to uniquely determine the position of the source within an enclosure, in practice it is impossible to obtain neither RIR nor RTF without knowing the positions of the sound source and the microphone within the room beforehand.

4.1.3. Steered Response Power

Steered Response Power (SRP) and SRP with Phase Transform (SRP-PHAT) vectors can be considered the middle ground between the trivial acoustic features like TDoA and ideal features, like RIR or RTF. SRP-PHAT are obtainable in real world, are relatively high-dimensional and contain information about sound reflections within the room.

4.1.4. Properties of acoustic features

The most important property of all acoustic features in this investigation is the spatial smoothness of feature space. In other words, acoustic features are similar to each other for sound source positions that are close together.

In our investigation, we use the SRP-PHAT spatial vectors as acoustic features.

4.2. Acoustic feature acquisition

Acoustic features were obtained within an acoustic enclosure using a single sound source, z coordinate was fixed at height $\mathrm{m_s}$. $\mathrm{N_M}$ circular microphone arrays were used for acoustic signal acquisition, each with $\mathrm{N_m}$ microphone elements and radius $\mathrm{m_M}$. Planes of the microphone arrays were parallel to the ground (normals of the circles coincided with the z axis of the enclosure). The both arrays were held at a fixed height $\mathrm{m_M}$. Signals of the microphones are recorded at sampling frequency f_s and resolution Q.

Audio signals obtained from the microphone arrays are split into frames of duration T_fr seconds to obtain N_fr frames. For each audio frame $j \in N_\mathrm{fr}$ and for each microphone array $i \in \mathrm{N_M}$, a time-frequency representation is calculated with $\mathrm{N_{FFT}}$ FFT points and $\mathrm{N_{hop}}$ an SRP-PHAT spatial spectrum $\mathbf{S}_{\text{SRP-PHAT}(j,i,(x,y,z))}$ is obtained. SRP-PHAT spectra of all arrays are then concatenated per frame to obtain the acoustic feature.

If the audio recording has an associated location label (known coordinates), a frame is assigned the position label $\vec{s}^{(x,y,z)}_{i,(x,y,z)}$.

It is considered that the sound source might not be active a all times, and that the signal is non-stationary (in case of speech signal). Thus, in case of an audio frame where the source is not active, the DoA of a sound source can not be determined, and the acoustic feature is considered to contain only noise. Such frames are to be discarded. For the selection of the audio frames in which the acoustic feature is usable, a thresholding algorithm was used. For each of the frame, a metric $p_{i,j}$ was calculated for and compared to the scaled mean of the metric of all obtained frames $\mathrm{k}_p \sum_{j \in N_\mathrm{fr}} p_j$, where k_p is the scaling coefficient used to control the threshold value. The metric used to evaluate the fitness of the acoustic feature of the particular audio frame are:

1. Root-mean-square value of the SRP-PHAT spectrum, $p_{\mathrm{RMS},i,j}$.
2. Crest factor of the SRP-PHAT spectrum, $p_{\mathrm{CF},i,j}$.

4.3. Acoustic manifold embedding learning

Manifold embedding can be learned using a Nonlinear Dimensionality Reduction (NLDR) algorithm, such as isometric mapping (ISOMAP), t-distributed stochastic neighbor embedding (t-SNE) or localli linear embedding (LLE).

4.3.1. ISOMAP embedding

SRP-PHAT features are embedded into $\mathrm{D_{ISO}}$-dimensional embedded space using ISOMAP, with $\mathrm{k_{ISO}}$ nearest neighbors considered. In the resulting embedding, SRP-PHAT features are grouped by similarity, while at the same time preserving the spatial structure of the high-dimensional (?) space

kaip tinkamai aprasyti FFT parametrus?

Describe the SRP-PHAT acquisition

is speech sign stationary or not? how to determine

scaled mean?

maybe that's energy of a frame?

kaip prasyti formule SRP spektrui? ten viena dimens yra kampas - integralas ar suma pagal kampa?

4.4. Graph dataset

4.4.1. Dataset preprocessing

In order to train the ANN with graph regularization, the dataset must be preprocessed: for each sample, alongside the main input feature, neighbor input features and their weights must be introduced. This is don by first determining the k_{ISO} nearest neighbors of a particular sample in the feature embedded space and then appending those features as well as their weight coefficients to the training sample.

4.4.2. Constitution of the dataset

The dataset for training the neural network is comprised of N source position and SRP-PHAT feature vector pairs (see Figure **??**).

In the embedded space, Euclidean distances are calculated between every point. The distances beteween each data sample constitute the distance matrix, which is in turn used to calculate the affinity matrix.

4.4.3. Affinity matrix calculation

Affinity matrix $\mathbf{A}$ is calculated by subtracting the distance matrix $\mathbf{D}$ from 1: $\mathbf{A} = 1 - \mathbf{D}$. The distance matrix contains the Euclidean distances between each sample in the ISOMAP embedded space:

$$\mathbf{D} = (d_{ij}); \tag{2}$$

$$d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2^2 \tag{3}$$

here $\mathbf{p}_i = (\alpha_i, \beta_i)$ is the point coordinate vector in the embedded space (in case of $N_{ISO} = 2$), α and β are the Cartesian coordinates in the embedded space.

Neighbor weights are inversely proportional to the Euclidean distances between the main feature and the neighbor features in the low-dimensional embedded space.

4.4.4. Neighbor samples and neighbor sample weights

For the training of the GRNN, each training sample must contain the main SRP-PHAT feature and k_g neighbor SRP-PHAT features (used for calculating the graph loss). Additionally, each neighbor feature is associated with its weight, which is the corresponding element in the affinity matrix. To obtain the k_g neighbors of each sample, each row of the affinity matrix is thresholded so that only the k_g highest-valued elements remain their value, while other row elements are set to zero. The dataset is then expanded so that each sample now has associated neighbor SRP-PHAT features (indices of which are the non-zero elements in the rows of the affinity matrix).

4.4.5. Training/testing dataset split

The preprocessed dataset is split into training and testing subsets so that $N_{\text{tr.}} = N - N_{\text{ts.}}$, and $N_{\text{ts.}} = 100$. $N_{\text{tr.}}$ is the size of the training subset, in samples, and $N_{\text{ts.}}$ is the size of the testing subset, in samples.

4.4.6. Labeled/unlabeled sample marking

For the training dataset, a flag m denoting wether the sample is labeled or unlabeled is introduced. This flag holds value of either "True" of "False" (1 or 0). Content of this field is interpreted by the GRNN during the calculation of the loss function. Effectively, the supervised loss component is multiplied by the flag. In case of an unlabeled sample, the supervised loss is ignored, and only

the graph loss is considered. In real-world scenario, GRNN expects all fields, including the target feature (the label, the coordinates of the source) to be passed during training. In case of the unlabeled sample (whether during the training phase or during the prediction phase), the supervised loss is not calculated, the label is ignored, and thus it can be set to random values or to zero.

4.4.7. Labeled samples repetition

We wish to train the GRNN using as few as possible labeled samples. It was found that the network is trained more effectively when the labeled samples are introduced more times (more often) than the unlabeled samples. It might be called "dataset balancing" []. Labeled samples (those with $m = 1$) are repeated N_R times ($N_R \in \{1, \ldots, 199\}$) and appended to the training data subset.

4.5. Graph-Regularized Neural Network

In our proposition, a neural network that is trained considering not only the labeled samples, but also neighboring labeled and unlabeled samples.

4.5.1. Neural network

Any neural network can be converted to graph-regularized neural network (GRNN) by introducing additional inputs for neighboring features as well as modifying the loss function to accommodate the graph loss.

A general architecture (one of possibilities) of a GRNN is provided in Figure **??**. In this figure, dotted lines encompasses the input vectors. Dahsed lines inside the GRNN block denote prediction (a forward pass). The loss function is given by $L = m(\hat{y}_0 - y) + \sum_{i \in k_g} a_i(\hat{y}_0 - \hat{y}_1)$. The loss function is discussed further in more detail.

4.5.2. Architecture

Apart from the introduction of additional inputs (neighbor features, weights and flags), the actual neural network is just a multilayer perceptron. During prediction phase, only the main input contributes to the prediction.

In this experiment, a simple multilayer perceptron architecture was used. It contained these layers:

1. A 1080-dimensional input layer (to accept a concatenated SRP-PHAT feature using $N_M = 3$ microphone arrays, each covering 360° azimuth with 1° resolution).
2. A fully-connected layer with 10 units and linear activations.
3. A fully-connected layer with 31 units and ReLU activations.
4. A 3-dimensional output layer with linear activations.

This architecture was the found during previously performed hyperparameter optimization.

4.5.3. Loss function

Nearby source positions produce similar acoustic features. Therefore, the predicted source positions for the nearby acoustic features should also be similar If they are similar, the graph loss is small. If they are not similar, we need to penalize the predictor with a large graph loss

Loss function

The loss function used for the GRNN training is comprised of two parts: the supervised loss (the difference between the ground truth label and the predicted label) and the graph loss (the difference bewteen the main input feature label prediction and the weighted sum of neighbor input features label predictions). It can be expressed as

$$L = \mu m \sum_{i \in N_b} (\hat{y}_i - y_i)^2 + (1 - \mu m) \sum_{i \in N_b} \sum_{j \in k_g} a_{ij} (\hat{y}_i - \hat{y}_j)^2 \tag{4}$$

here N_b – number of samples in one training batch, k_g – size of the neighborhood, a_{ij} is the neighbor weight, equal to the corresponding element in the affinity matrix.

5. Conclusions

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

6. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing–original draft preparation, X.X.; writing–review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text. For example, explanations of experimental details that would disrupt the flow of the main text, but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with 'A', e.g., Figure A1, Figure A2, etc.

References

1. Marti, A.; Cobos, M.; Aguilera, E.; Lopez, J.J. Speaker Localization and Detection in Videoconferencing Environments Using a Modified SRP-PHAT Algorithm. p. 8.
2. Laufer-Goldshtein, B.; Talmon, R.; Gannot, S. Semi-Supervised Sound Source Localization Based on Manifold Regularization. *24*, 1393–1407. doi:10.1109/TASLP.2016.2555085.
3. He, W.; Motlicek, P.; Odobez, J. Deep Neural Networks for Multiple Speaker Detection and Localization. 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 74–79. doi:10.1109/ICRA.2018.8461267.
4. He, W.; Motlicek, P.; Odobez, J.M. Adaptation of Multiple Sound Source Localization Neural Networks with Weak Supervision and Domain-Adversarial Training. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 770–774. doi:10.1109/ICASSP.2019.8682655.

Sample Availability: Samples of the compounds ..... are available from the authors.