Journal of Siberian Federal University. Engineering & Technologies 2024 17(3): 380–398
~ ~ ~

EDN: TSIHQW

УДК 004.032.26, 004.048

Analysis of Approaches and Methods to Acoustic Sources Localization

Ghiath M. Shahoud* and Evgeny D. Agafonov
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 03.02.2024, received in revised form 20.03.2024, accepted 05.04.2023

Abstract. This article provides an overview of traditional methods to acoustic sources localization based on signal processing, as well as modern methods based on the use of deep neural networks. The advantages and disadvantages of the above methods are analyzed and discussed. Although some traditional methods can adapt to observed signals, they all depend on assumptions made about the nature of the environment, the properties of the signals, etc. Deep learning models do not explicitly require any of these assumptions, but instead efficiently adapt to the training data provided. However, this is also a major disadvantage of modern methods, as they are less generalizable and less versatile than traditional methods. A justification is given for the need to develop new localization methods, as well as the integration of traditional and intelligent modern localization methods to combine the advantages of each of these groups of methods.

Keywords: acoustic sources localization, signal processing, deep neural networks, training data.

Citation: Shahoud Gh. M., Agafonov E. D. Analysis of approaches and methods to acoustic sources localization. J. Sib. Fed. Univ. Eng. & Technol., 2024, 17(3), 380–398. EDN: TSIHQW

* Corresponding author E-mail address: ghiathlovealaa@gmail.com; eagafonov@sfu-kras.ru

Анализ подходов и методов локализации акустических источников

Д. М. Шаход, Е. Д. Агафонов
Сибирский федеральный университет
Российская Федерация, Красноярск

Аннотация. В данной статье представлен обзор традиционных методов локализации акустических источников, основанных на обработке сигналов, а также современных методов, основанных на применении глубоких нейронных сетей. Проанализированы и рассмотрены преимущества и недостатки приведенных методов. Несмотря на то что некоторые традиционные методы могут адаптироваться к наблюдаемым сигналам, все они зависят от принятых предположений и допущений о характере среды, о свойствах сигналов и т.д. Модели глубокого обучения явно не требуют ни одного из этих предположений, а вместо этого эффективно адаптируются к предоставленным обучающим данным. Однако это также является основным недостатком современных методов, поскольку они менее способны к обобщению и менее универсальны, чем традиционные методы. Дано обоснование необходимости развития новых методов локализации, а также интеграции традиционных и современных интеллектуальных методов локализации для объединения преимуществ каждого из этих групп методов.

Ключевые слова: локализация акустических источников, обработка сигналов, глубокие нейронные сети, обучающие данные.

Цитирование: Шаход Д. М. Анализ подходов и методов локализации акустических источников / Д. М. Шаход, Е. Д. Агафонов // Журн. Сиб. федер. ун-та. Техника и технологии, 2024, 17(3). С. 380–398. EDN: TSIHQW

Введение

Необходимость решения задач, связанных с определением местоположения (локализации) источников акустических сигналов (Sound Source Localization, SSL) или объектов, отражающих акустические сигналы, возникает во многих сферах и приложениях техники и технологии. В частности, локализация акустических источников широко применяется при автоматическом слежении за камерой для телеконференций, взаимодействии человека с роботом, распознавании речи на расстоянии, мониторинге и громкой связи. Также задачи локализации имеют исключительную важность в геофизике и в технике военного назначения.

Для регистрации акустических сигналов с дальнейшим анализом их характеристик применяются микрофонные решетки, которые состоят из набора микрофонов различной геометрии, расположенных в пространстве для получения пространственной информации об акустическом источнике. Пространственно-временная информация, полученная с микрофонной решетки, может использоваться для оценки различных параметров источника или извлечения предполагаемого исходного сигнала. Микрофонные решетки применяются для локализации и отслеживания нескольких акустических источников [1], обнаружения и классификации акустических событий [2], распознавания говорящего [3], снижения шума [4], подавления акустического эха [5]. Микрофонную решетку также можно применять при локализации движущихся объектов, таких как дроны [6], для их обнаружения и классификации и, таким образом, определения степени их опасности.

Немаловажную роль в локализации акустических сигналов играют методы и алгоритмы анализа и обработки сигналов. Первоначально задача локализации акустических источников решалась с помощью традиционных методов обработки сигналов, таких как Time Difference of Arrival (TDOA) [7], Delay-And-Sum beamformer (DAS) [8], Multiple Signal Classification (MUSIC) [9] и Generalized cross-correlation – phase transform (GCC-PHAT) [10]. Однако эти методы имеют недостатки, связанные со сложностью акустических характеристик окружающей среды. В последние годы с появлением и развитием методов глубокого обучения, таких как Convolutional Neural Networks (CNN) [11], Recurrent Neural Network (RNN) [12], Convolutional Recurrent Neural Network (CRNN) [13] и Residual neural networks (ResNet) [14] и их широким использованием в области акустических приложений, был намечен новый вектор развития направления локализации источников. Исследования доказали эффективность методов глубокого обучения в решении рассматриваемых проблем, с которыми не справляются традиционные методы. Особенно заметные результаты достигнуты при анализе сигналов с наличием существенного шума и реверберации [13]. Таким образом, современные исследования в области локализации акустических источников направлены на разработку методов глубокого обучения с использованием сигналов микрофонной решетки.

Целью данной статьи является обзор и анализ подходов и методов локализации акустических сигналов, в том числе с использованием глубокого обучения, а также рекомендации по интеграции традиционных и современных методов для выработки универсального и эффективного подхода решения проблемы локализации.

1. Микрофонные решетки

Микрофонная решетка – это несколько микрофонов, расположенных в разных пространственных точках, которые используются для получения дополнительной доступной пространственной информации [15]. Сигнал, полученный микрофонной решеткой, имеет разнесение в пространственной области с учетом временных задержек между микрофонами решетки. С другой стороны, пространственная дискретизация позволяет спроектировать пространственный фильтр, пропускающий источники с определенных направлений и отклоняющий источники с других направлений [16]. Этот метод пространственной фильтрации также называется формированием луча.

Пространственно-временная информация может оказывать влияние на эффективность операций обработки речи, локализацию акустического источника и реальных приложениях.

Микрофонные решетки могут иметь различную геометрию для приема акустического сигнала. Широко используются линейные, плоские и сферические микрофонные решетки. Хотя линейная микрофонная решетка проста по структуре и обработке, она ограничена неоднозначностью. Плоские решетки преодолевают неоднозначность «перед-зад», однако не справляются с неоднозначностью «верх-низ». Сферическая микрофонная решетка позволяет локализовать источники в любой точке пространства без пространственной неоднозначности [16].

Формы акустических волн, распространяющихся в пространстве, связаны с типом поля (ближнее или дальнее).

Ближнее поле определяется как диапазон, в котором микрофоны расположены близко к акустическому источнику, на расстоянии меньше длины волны, соответствующей самой низ-

кой частоте источника сигнала. В этом диапазоне акустический сигнал распространяется как сферическая волна [17].

Дальнее поле определяется как диапазон, в котором акустический источник находится на расстоянии от всех микрофонов решетки, превышающем длину волны, и простирается до бесконечности, принимая во внимание, что расстояние между микрофонами в решетке достаточно мало по сравнению с минимальной длиной волны источника сигнала [17]. В этом диапазоне акустический сигнал распространяется как плоская волна.

1.1. Линейные микрофонные решетки

Равномерная линейная решетка (Uniform Linear Array, ULA) представляет собой простейшую конфигурацию решетки. На рис. 1 показана ULA с M микрофонами, равномерно расположенными по оси X. Расстояние между двумя последовательными микрофонами равно d.

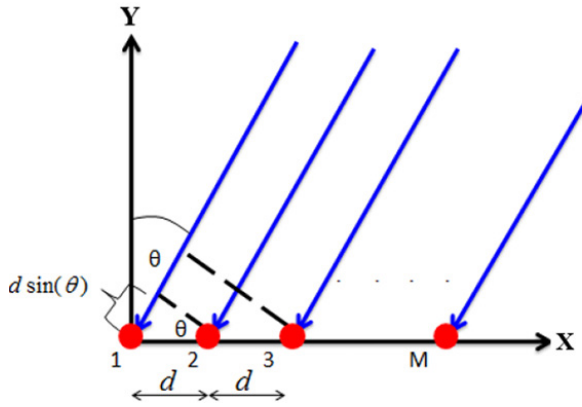

Рис. 1. Геометрия однородной линейной решетки

Fig. 1. Geometry of a uniform linear array

Предполагаем, что источник находится в дальнем поле и акустическая волна плоская. Источник в дальнем поле падает на решетку под азимутальным углом θ. Дополнительное расстояние, проходимое волновым фронтом между двумя последовательными микрофонами, равно *dsin*(θ). Можно заметить, что азимут источника по отношению к прямому направлению равен углу θ. Для геометрии горизонтальной плоскости опорным направлением является прямое направление (ось Y), а θ – азимут внутри плоскости. ULA может локализовать источники только в своей плоскости с азимутом в диапазоне $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$ (Направление оси Y соответствует θ = 0, а направление оси X соответствует θ = π/2). ULA может использоваться для оценки угла акустического источника, однако не подходит для определения, с какой стороны ось X. Это называется неоднозначность «перед-зад».

1.2. Плоские микрофонные решетки

Наиболее распространенными плоскими микрофонными решетками являются равномерная круговая решетка (Uniform Circular Array, UCA) и ортогональная решетка (*Orthogonal Array, OA*).

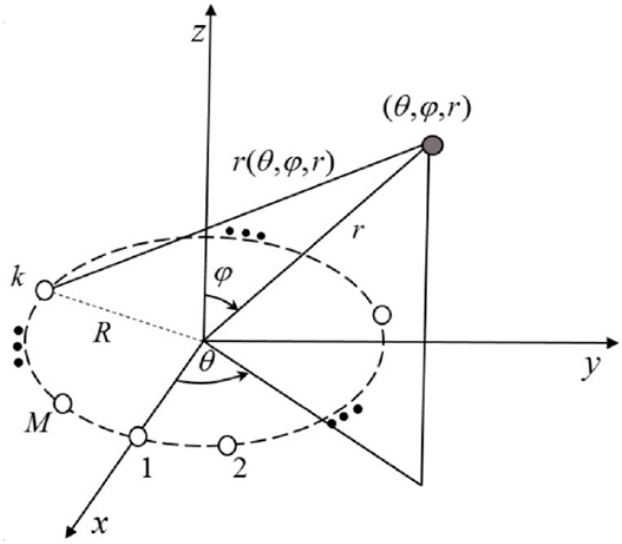

Рис. 2. Геометрия однородной круговой решетки

Fig. 2. Geometry of a uniform circular array

В равномерной круговой решетке микрофоны расположены равномерно по окружности, как показано на рис. 2. UCA может локализовать источники с любым азимутом, т.е. $\theta \in [0,2\pi]$ и углом места (элевация) φ от 0 до $\pi/2$.

Хотя UCA не справляются с неоднозначностью «перед-зад», они ограничены неоднозначностью «верх-низ» [16]. Еще одним преимуществом является то, что UCA намного компактнее, чем ULA, при том же количестве микрофонов и условиях пространственного перекрытия.

2. Традиционные методы локализации акустических сигналов

2.1. Разность во времени приема сигнала (Time Difference of Arrival, TDOA)

Основная идея TDOA заключается в использовании разницы во времени приема акустических сигналов между микрофонами решетки. Поскольку акустический сигнал распространяется как плоская волна, наблюдается задержка между различными микрофонами решетки. Временные задержки τ_{ij} между *i*-m и *j*-m микрофонами можно выразить как [18]:

$$\tau_{ij} = \frac{d_{ij}}{c} = \frac{\|x_s - r_i\| - \|x_s - r_j\|}{c}, \quad (1)$$

где d_{ij} – разница расстояний от источника до *i*-го и *j*-го микрофонов, r_i и r_j – местоположения *i*-го и *j*-го микрофонов, c – скорость звука, а x_s – местоположение источника. Выражение (2) для двумерного пространства с использованием местоположений микрофонов, обозначенных как (x_i, y_i) и (x_j, y_j), и источника (x_s, y_s) имеет вид [19]:

$$d_{ij} = \sqrt{(x_s - x_i)^2 + (y_s - y_j)^2} - \sqrt{(x_s - x_j)^2 + (y_s - y_j)^2}. \quad (2)$$

Временную задержку можно рассчитать путем сравнения двух сигналов с использованием функции кросс-корреляции. Основная идея состоит в том, чтобы сравнить сходство характеристик двух сигналов. Функция кросс-корреляции дает пик, представляющий временную задержку [20]. Для достижения более высокой производительности кросс-корреляция обычно рассчитывается в частотной области. Сигналы преобразуются в частотную область с использованием метода преобразования Фурье, такого как дискретное преобразование Фурье (Discrete-Time Fourier Transform, DTFT) или быстрое преобразование Фурье (Fast Fourier Transform, FFT) [18], [7].

Из уравнений (1) и (2), рассчитав временные задержки между микрофонами и зная расположение микрофонов в решетке, можно определить местоположение акустического источника (x_s, y_s).

При оценке направления приема (Direction of Arrival, DOA), также известного как пеленг (азимут), обычно используется модель распространения в дальнем поле [21], [22]. На рис. 3a показан пример линейной микрофонной решетки, использующей модель дальнего поля. DOA можно выразить с использованием полярных координат как:

$$\theta = \arctan\left(\frac{x_s - x_i}{y_s - y_i}\right), \tag{3}$$

где (x_s, y_s) – местоположение источника, а (x_i, y_i) – местоположение i-го микрофона.

Преимущество TDOA состоит в том, что не нужно знать время приема сигнала и она позволяет осуществлять распределенную обработку, а время вычислений можно считать относительно небольшим [18]. Недостатки TDOA включают чувствительность к проблемам синхронизации между микрофонными решетками и недостаточную устойчивость к внешним помехам, таким как шум и реверберация [7]. Предполагая, как показано на рис. 3b, что сенсорные узлы

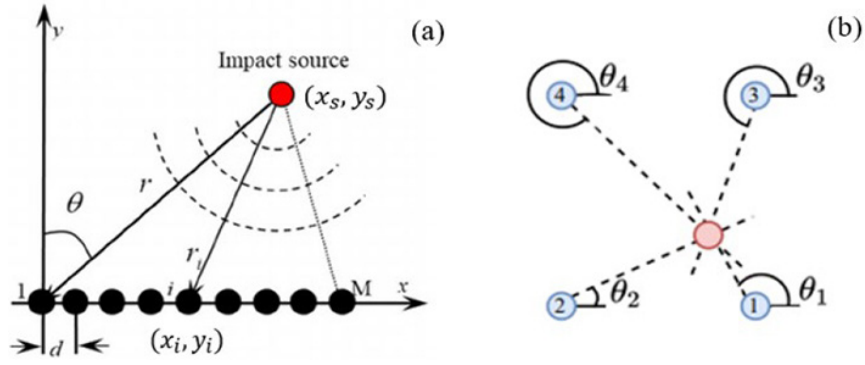

Рис. 3. a) иллюстрация разницы в расстоянии, которое сигнал проходит от источника до обоих микрофонов 1 и i в линейной решетке в предположении дальнего поля; b) иллюстрация принципа поиска акустического источника (красная точка) с использованием пересекающихся DOAs, оцененных с использованием нескольких микрофонных решеток (каждая синяя точка представляет микрофонную решетку)

Fig. 3. a) Illustration of the difference in distance the signal travels from the source to both microphones 1 and i in a linear array, assuming far-field. b) Illustration of the principle of acoustic source search (red dot) using intersecting DOAs estimated using multiple microphone arrays (each blue dot represents a microphone array)

имеют несколько решеток микрофонов, вычисления временных задержек могут выполняться отдельно каждым узлом [18]. Это является предпочтительным вариантом, поскольку тогда центральному процессору нужно будет только объединить полученные измерения TDOA для оценки положения источника. Кроме того, это делает систему менее чувствительной к проблемам синхронизации между узлами [18].

2.2. Формирование луча (Beamforming)

Формирование луча – это метод акустической визуализации. Основная идея акустической визуализации состоит в том, чтобы отобразить картину окружающей среды с использованием данных, собранных с помощью микрофонных решеток, и модели распространения звука [23]. Используя полученные данные и параметры модели распространения звука, можно получить желаемые характеристики, которыми в случае формирования луча являются направленность принимаемого сигнала. Как правило, алгоритмы формирования луча оценивают направление принимаемых сигналов путем усиления сигналов с определенных направлений и ослабления сигналов с других направлений [8].

Самым простым, но широко используемым методом формирования луча является формирователь луча с задержкой и суммированием DAS [21]. Алгоритм DAS основан на задержке принимаемых сигналов на каждом микрофоне, что делается для компенсации относительных задержек времени приема сигналов [8]. На рис. 4 показана базовая иллюстрация алгоритма DAS, где сигнал регистрируется тремя микрофонами, затем сигналы микрофонов сдвигаются по времени в зависимости от расположения каждого микрофона и далее суммируются для формирования выходного сигнала. Общая идея состоит в том, что задержанные сигналы синхронизируются, если исходят в одном направлении. Сумма выравнивающих сигналов приведет к усиленному выходному сигналу, соответствующему оценке направления [21].

Выходной сигнал алгоритма DAS можно выразить во временной области как [8]:

$$B\left(\overrightarrow{x_p}, t\right) = \frac{1}{M}\sum_{m=1}^{M} w_m A_m\left(\overrightarrow{x_p}, \overrightarrow{x_m}\right) P_m\left(t - \tau_{pm}\right), \tag{4}$$

где M – количество микрофонов, w_m – весовой коэффициент на m-м микрофоне, A_m – коэффициент масштабирования амплитуды, P_m – сигнал давления на m-м микрофоне, τ_{pm} – временная задержка между сигналами опорного микрофона и m-м микрофоном, соответствующая местоположению источника $\overrightarrow{x_p}$ и $\overrightarrow{x_m}$– местоположение m-го микрофона.

Чтобы выразить выходной сигнал в частотной области, вместо временной задержки используется фазовая задержка, которую можно выразить как [8]:

$$B\left(\overrightarrow{x_p}, w_k\right) = \frac{1}{M}\sum_{m=1}^{M} w_m A_m\left(\overrightarrow{x_p}, \overrightarrow{x_m}\right) P_m(w_k) e^{jw_k\tau_{pm}}, \tag{5}$$

где w_k – угловая частота, т.е. $w = 2\pi f$ (f – частота акустического сигнала), а $e^{jw_k\tau_{pm}}$– фазовая задержка.

Недостатком метода DAS является относительно низкое пространственное разрешение. Это, в свою очередь, может привести к появлению так называемых мнимых образов, что означает, что алгоритм формирования луча выводит дополнительные, несуществующие источни-

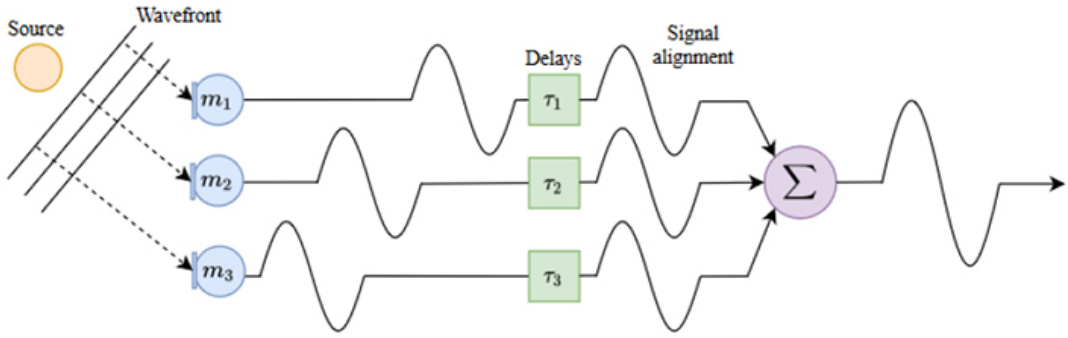

Рис. 4. Основная идея функционирования формирователя луча DAS: каждый микрофон улавливает акустический сигнал, полученные сигналы задерживаются и суммируются. Если сигналы исходят из одного и того же направления, сумма выравнивающих сигналов создаст усиленный выходной сигнал

Fig. 4. The basic idea of how a beamformer DAS works: each microphone picks up an acoustic signal, the resulting signals are delayed and summed. If the signals come from the same direction, the sum of the alignment signals will create an amplified output signal

ки. По сравнению с TDOA методы на основе формирования луча устойчивы к реверберации и помехам, таким как шум. Однако обычно считается, что они обеспечивают решения с плохим пространственным разрешением, а в случае больших микрофонных решеток вычислительные затраты довольно высоки.

2.3. Multiple Signal Classification (MUSIC)

MUSIC – это алгоритм пространственной спектральной оценки, основанный на подпространствах, который может оценивать направление одного или нескольких акустических источников.

Предполагая, что сигналы и шум целевого источника некоррелированы, метод MUSIC применяет разложение собственных значений (Eigenvalue Decomposition, EVD) для оценки подпространств сигнала и шума [24].

Как видно на рис. 5, однородная линейная решетка состоит из N микрофонов. A(θ, *w*) определяется как вектор управления (steering vector) [24], содержащий фазовые задержки, следующим образом:

$$A(\theta, w) = [1 \ \exp(-jw\tau_1) \dots \dots \exp(-jw\tau_{N-1})]^T, \quad (6)$$

$$\tau_i = \frac{iDsin(\theta)}{c}; i = 1. \dots . N - 1, \quad (7)$$

где D – расстояние между двумя последовательными микрофонами, θ – угол азимута. *c* – скорость звука и τ_i – временная задержка между *i*-м микрофоном и опорным микрофоном с номером 0, т.е. $\tau_0 = 0$.

Пусть вектор *x*(*n*) представляет собой набор N микрофонных сигналов во временной выборке *n* и формулируется следующим образом:

$$x(n) = [x_0(n) \ x_1(n) \dots \dots x_{N-1}(n)]^T, \quad (8)$$

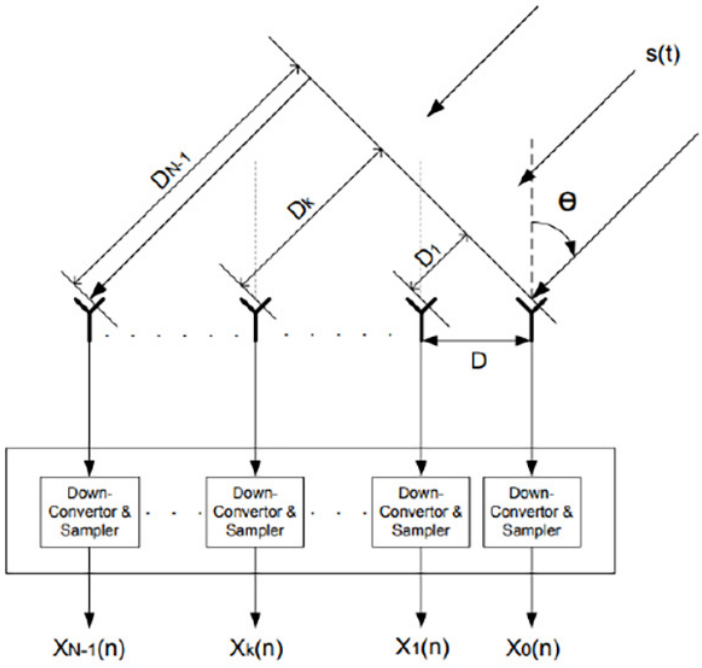

Рис. 5. Схема распределения акустического источника и решетки микрофонов

Fig. 5. Diagram for distribution of an acoustic source and microphone array

представление вектора $x(n)$ в частотной области можно получить, приняв преобразование Фурье, следующим образом:

$$X(w) = S(w)A(\theta, w) + V(w), \tag{9}$$

где $S(n)$ – преобразование Фурье сигнала источника $s(t)$ и $V(w)$ – набор преобразований Фурье шума на каждом микрофоне, формулируется в виде вектора следующим образом:

$$V(w) = [V_0(w)\, V_2(w)\ \ldots\ldots V_{N-1}(w)\ ]^T, \tag{10}$$

предполагается, что сигнал источника представляет собой узкополосный сигнал с центральной частотой w_0, поэтому выражение (9) можно записать в следующем виде:

$$X(w_0) = S(w_0)A(\theta, w_0) + V(w_0). \tag{11}$$

Матрица пространственной корреляции вектора $X(w_0)$ представляет собой матрицу размера N × N, определяемую выражением:

$$R = XX^H. \tag{12}$$

Основная идея алгоритма MUSIC состоит в том, чтобы получить подпространство сигнала и подпространство шума посредством разложения по собственным значениям матрицы R, а затем оценить параметр сигнала, используя ортогональность двух пространств [9]. Для описания ортогональности между подпространством сигнала и подпространством шума используется пространственный спектр, который можно рассчитать по формуле (13):

$$P_{MUSIC}(\theta) = \frac{1}{A^H(\theta, w_0) U_N U_N^H A(\theta, w_0)}, \quad (13)$$

где U_N обозначает подпространство шума, охватываемое собственным вектором, соответствующим наименьшему собственному значению [9].

На основании формулы (13) $A(\theta, w_0)$ управляется для сканирования направлений источников в пространстве. Пиковая точка пространственного спектра $P_{MUSIC}(\theta)$ соответствует направлению акустического источника.

Недостатком метода MUSIC является то, что он применяется к узкополосным сигналам, поэтому узкополосные сигналы с определенной центральной частотой должны быть извлечены из широкополосных волн, генерируемых каждым источником звука, с использованием вейвлет-преобразования. Кроме того, конкретная центральная частота получается после тщательного анализа исходного сигнала, который занимает много времени. Кроме того, методы подпространства устойчивы к шуму и могут давать очень точные оценки, но они чувствительны к реверберации.

Авторы в [25] представили улучшенный подход, который сочетает в себе алгоритм оптимизированной ансамблевой эмпирической модовой декомпозиции (Ensemble Empirical Mode Decomposition, EEMD) и MUSIC для локализации акустического источника в реальном времени. Во-первых, исходный сигнал с неизвестным азимутом регистрируется с помощью линейной решетки микрофонов. Во-вторых, введено быстрое преобразование Гильберта Хуанга (Hilbert Huang Transform, HHT) с EEMD для извлечения функций внутреннего режима (Intrinsic Mode Functions, IMFs) из исходных сигналов. Затем все IMFs во всей частотной области напрямую используются в качестве входного вектора модели MUSIC отдельно для определения местоположения акустического источника. Результаты показали, что использование оптимизированных EEMD и MUSIC подходит для локализации источника в реальном времени.

2.4. Обобщенная взаимная корреляция – фазовое преобразование (Generalized cross-correlation – phase transform, GCC-PHAT)

Алгоритм локализации с помощью обобщенной функции взаимной корреляции с функцией преобразования фазы, разработанный в 1976 году Кнаппом и Картером [10], может уменьшить эффекты автокорреляции сигнала и сделать систему более устойчивой к реверберации. Однако GCC-PHAT может быть неприменим к решетке небольшого размера.

GCC-PHAT часто используется для расчета TDOA в одновременных акустических сигналах в средах с умеренной реверберацией с использованием классических методов обработки сигналов.

GCC-PHAT – один из наиболее часто используемых методов при работе с решеткой из двух микрофонов [10]. Подход GCC был распространен на решетки с более чем двумя микрофонами. Продемонстрировано, в частности, что локализацию можно улучшить, воспользовавшись преимуществами нескольких пар микрофонов [26]. GCC-PHAT рассчитывается как обратное преобразование Фурье перекрестного спектра мощности (Cross-Power Spectrum, CPS) между сигналами двух микрофонов.

Пусть x_i и x_j будут двумя сигналами, CPS определяется как:

$$CPS = \frac{X_i(f)X_j^*(f)}{\left|X_i(f)X_j^*(f)\right|}, \tag{14}$$

где $X_i(f)$ и $X_j(f)$ – N-точечные дискретные преобразования Фурье двух сигналов, символ (*) означает комплексно-сопряженный.

Обратное преобразование Фурье для CPS определяется как:

$$\hat{R}_{PHAT}(\tau) = \sum_{f=0}^{F-1} \frac{X_i(f)X_j^*(f)}{\left|X_i(f)X_j^*(f)\right|} e^{j2\pi\left(\frac{f\tau}{N}\right)}. \tag{15}$$

Оценка временной задержки между двумя микрофонами определяется путем нахождения значения τ, которое максимизирует функцию $\hat{R}_{PHAT}(\tau)$:

$$\hat{\tau}_{PHAT}(i,j) = \underset{\tau}{\operatorname{argmax}}\left(\hat{R}_{PHAT}(\tau)\right). \tag{16}$$

3. Современные методы локализации сигналов

Способность методов глубокого обучения (Deep Learning, DL) заменять традиционные методы, основанные на модели сигнала/канала, и методы обработки сигналов (Signal Processing, SP), или, по крайней мере, часть из них, поскольку модуль извлечения признаков может быть основан на традиционной обработке – делает их привлекательными для решения таких проблем, как SSL.

Таким образом, применение глубоких нейронных сетей (Deep Neural Networks, DNNs) предлагается в литературе для решения проблемы SSL [27]. Проектирование DNN для конкретного приложения часто требует исследования (и, возможно, объединения) различных архитектур и настройки их гиперпараметров. Так было с SSL в течение последнего десятилетия, и эволюция методов SSL на основе DL следовала за общей эволюцией DNN в сторону все более и более сложных архитектур или новых эффективных моделей, принятых сообществами DL и SP в целом [27]. Другими словами, архитектуры DNN, используемые в SSL, часто унаследованы от других работ в других (связанных или более отдаленных) областях просто потому, что было показано, что они хорошо работают с акустическими сигналами или другими типами сигналов. По той же методике часто комбинируются (параллельно и/или последовательно) разные модели.

3.1. Сверточная нейронная сеть (Convolutional Neural Networks, CNN)

CNN широко используется для распознавания образов и успешно применяется для решения различных задач, таких как классификация изображений, обработка естественного языка (Natural Language Processing, NLP), автоматическое распознавание речи и SSL.

В [11] CNN использовалась для оценки пространственного местоположения и классификации сигнала акустического источника, расположенного в 8 разных местах пространства, на два класса (речевой или музыкальный источник). Сеть обучалась на акустических сигналах, полученных из двух решеток микрофонов разных размеров, каждая решетка содержит 8 микрофонов. Предложенная модель показала, что она может давать хорошие результаты и адапти-

роваться к различным конфигурациям (разные размеры микрофонной решетки), однако автор указал на актуальную проблему такой модели – устойчивость сети к изменению местоположения источника. То есть не удается получить удовлетворительные результаты при изменении местоположения источников от того, каким они были при обучении сети.

Также предлагались CNN для прогнозирования азимута одного или двух динамиков в реверберирующей среде. Входными признаками являлись многоканальные фазовые спектрограммы STFT. Предложенная модель показала превосходящую производительность по сравнению с методами GCC-PHAT и MUSIC. Также было продемонстрировано, что для получения более высокой производительности модели с использованием решетки из M микрофонов оптимальным количеством сверточных слоев для использования фазовых корреляций между соседними микрофонами является M – 1 [28].

Авторы в [29] сравнили 4-слойную MLP (Multi-Layer Perceptron) и 4-слойную CNN для решения задачи обнаружения и локализации нескольких говорящих. Результаты оценки показали одинаковую точность для обеих архитектур. Также показали, что предложенные методы существенно превосходят традиционные методы, основанные на пространственном спектре. Однако они ограничены выборками обучающих данных, которые вряд ли охватывают все возможные комбинации акустических источников. Следовательно, эти модели не могут быть обобщены на несколько акустических источников.

Более глубокая архитектура CNN, от 11 до 20 сверточных слоев, была предложена в [14]. CNN доказала устойчивость к шуму по сравнению с MUSIC, а также меньшее время обучения. Однако для обеспечения надежной работы в сложных условиях модель требует дополнительных вычислительных затрат на обучение.

Хотя большинство систем локализации направлены на оценку азимута или азимута и угла места, авторы в [30] исследовали оценку только угла места с использованием CNN с бинауральными входными признаками. Авторы в [31] применили CNN непосредственно к необработанным многоканальным сигналам для прогнозирования декартовых координат (x, y, z) одного статического или движущегося динамика. Эксперименты показали, что предложенная модель демонстрировала хорошую устойчивость к разному полу говорящего и разным размерам окна входных сигналов.

Авторы в [32] предложили использовать двумерную сверточную нейронную сеть с многозадачным обучением для надежной оценки количества источников (Number of Source, NOS) и направлений приема DOA на основе кратковременных пространственных псевдоспектров (short-time spatial pseudo-spectra), извлеченных с помощью алгоритма MUSIC, которые содержат полезную информацию о направлении из входных акустических сигналов. Они проверили свою модель на сигналах, содержащих до четырех источников, и показали хорошую производительность как в моделируемой, так и в реальной среде.

В литературе по оценке DOA есть несколько работ, посвященных использованию расширенных сверток в DNN. Расширенная свертка (Dilated Convolution) – это тип сверточного слоя, в котором ядро свертки шире классического, но вставляются нули, чтобы количество параметров оставалось прежним. Авторы в [33] демонстрируют, что включение расширенных сверток с постепенным увеличением коэффициентов расширения уменьшает оптимальное количество сверточных слоев их исходной архитектуры CNN. Это приводит

к созданию архитектуры с аналогичной производительностью и меньшими вычислительными затратами.

В [34] была предложена модель локализации с использованием небольшой микрофонной решетки, созданной из двух ортогональных решеток микрофонов. Также предлагалась улучшенная система извлечения признаков, основанная на оценке акустической интенсивности. Результаты моделирования и реальные эксперименты показали, что предлагаемый подход может достичь более высокого пространственного разрешения и превзойти свои современные аналоги, используя характеристики акустической интенсивности решетки небольшого размера в шумной и реверберирующей среде.

В [35] предлагалась модель на основе CNN для SSL в шумных и реверберирующих условиях с использованием микрофонных решеток небольшого размера. CNN была специально обучена для определения DOA акустического источника в этих условиях. Авторы создали три набора данных с помощью методов моделирования, которые учитывают различные акустические параметры, включая размер помещения, местоположение микрофонной решетки в помещении и расстояния между центром решетки и источником. Предложенная модель доказала свою способность обобщения по первым двум акустическим параметрам при условии расширения набора данных.

3.2. Рекуррентная нейронная сеть (Recurrent Neural Network, RNN)

RNNs – это нейронные сети, предназначенные для моделирования временных последовательностей данных. Конкретные типы RNN включают сети долгой краткосрочной памяти (Long Short Term Memory, LSTM) и управляемые рекуррентные блоки (Gated Recurrent Units, GRUs). Эти два типа RNN стали очень распространенными благодаря их способности обойти трудности обучения, с которыми сталкиваются обычные RNN, в частности проблему исчезновения и взрыва градиента.

Существует мало опубликованных работ по SSL, в которых используются только RNN, поскольку рекуррентные слои часто комбинируются со сверточными слоями. В [12] использовалась RNN для согласования прогнозов обнаружения акустических событий (Sound Event Detection, SED) и DOA (азимут и элевация), которые были получены отдельно для каждого возможного класса акустического события. В конечном итоге RNN использовалась для определения того, какой прогноз SED какой оценке DOA соответствует.

3.3. Сверточная рекуррентная нейронная сеть (Convolutional Recurrent Neural Network, CRNN)

CRNNs – это нейронные сети, которые содержат один или несколько сверточных слоев и один или несколько рекуррентных слоев. CRNNs регулярно используются для SSL. В последние годы из-за особых возможностей этих сетей CNNs оказались подходящими для извлечения соответствующих признаков для SSL, а рекуррентные слои хорошо справляются с проблемой интеграции информации с течением времени.

В [13] представлена модель с использованием блока сверточных слоев, блока слоев BGRU и слоя прямой связи. Эта модель оказалась способной локализовать и обнаружить несколько

акустических событий в трехмерном пространстве, даже если они перекрывались во времени, при условии, что они относились к разным классам (например, речь и автомобиль).

Предлагаемая CRNN в [13] была базовой системой для задачи 3-го конкурса «Обнаружение и классификация акустических сцен и событий, DCASE» в 2019 и 2020 годах. Таким образом, это послужило вдохновением для многих других работ, и многие модели были построены на основе системы, предложенной в [13], с различными модификациями и улучшениями. Например, авторы в [36] добавили гауссов шум к входным спектрограммам, чтобы научить сеть быть более устойчивой к шуму. Автор в [37] интегрировал несколько дополнительных сверточных слоев и заменил слои BGRU двунаправленными слоями LSTM. Признаки GCC-PHAT были добавлены в качестве входных [38]. Результаты показали, что использование этих признаков улучшило производительность предложенной модели в [13]. Авторы в [39] использовали аугментацию данных во время обучения и усредняли выходные данные сети для более стабильной оценки DOA. Авторы в [40] отправили входные признаки отдельно в разные ветви сверточных слоев, логарифмическую спектрограмму и признаки постоянного Q-преобразования, с одной стороны, и фазовые спектрограммы и признаки CPS, с другой стороны. В [41] авторы объединили логарифмическую спектрограмму и признаки GCC-PHAT и поместили их в две отдельные CRNNs для оценки SED и DOA.

Несколько CRNNs были обучены в [42]: одна для оценки NOS (до двух источников), другая для оценки DOA, предполагая один активный источник, третья для оценки DOA двух одновременно активных источников.

3.4. Остаточные нейронные сети (Residual neural networks, ResNet)

Остаточные нейронные сети были первоначально представлены автором в [43], который отметил, что проектирование очень глубоких сетей может привести к взрывному росту или “вымыванию” градиента из-за нелинейных функций активации, а также к ухудшению общей производительности. Остаточные связи предназначены для того, чтобы признаки могли обходить блок слоев параллельно традиционному процессу через этот блок слоев. Это обычно приводит к лучшему обучению.

Первое использование остаточной сети для SSL было предложено авторами в [14]. Эта сеть включает в себя три оставшихся блока, которые представляют собой стопки слоев, причем один из слоев имеет остаточные связи с другим уровнем, расположенным глубже в стеке. Каждый из этих блоков состоит из трех сверточных слоев, первый и последний из которых состоят из 1 × 1 фильтров, а средний слой состоит из 3 × 3 фильтров. Между входом и выходом каждого остаточного блока используется остаточные связи. Тот же тип остаточного блока использовался для SSL параллельно с классификацией звуков на речевые и неречевые [27], [44]. Использовалась серия одномерных сверточных слоев с несколькими остаточными связями для локализации одного источника непосредственно по многоканальному сигналу [27], [45].

4. Проблемы, связанные с эффективностью существующих методов

Традиционные методы, представленные алгоритмами TDOA, DAS, MUSIC и GCC-PHAT, имеют недостатки, которые включают:

1) чувствительность к проблемам синхронизации между микрофонными решетками и недостаточную устойчивость к внешним помехам, таким как шум и реверберация – TDOA,

2) относительно низкую пространственную точность – DAS,

3) устойчивы к шуму и могут давать очень точные оценки, но они чувствительны к реверберации, с другой стороны, их применение к широкополосным сигналам требуют больших вычислительных затрат – MUSIC,

4) устойчивы к шуму и могут быть обобщены, но могут быть неприменимы к решетке небольшого размера – GCC-PHAT.

Современные методы, представленные сетями CNN, RNN, CRNN и ResNet, способны превосходить традиционные методы за счет адаптации к большому объему данных, содержащих акустические характеристики, в сложной акустической среде при наличии шума и реверберации для точной оценки местоположения источника. Однако их недостатком является то, что они менее способны к обобщению, чем традиционные методы, при изменении настроек конфигурации.

Глубокая модель, разработанная и обученная для конкретной конфигурации (например, конкретной геометрии микрофонной решетки), не обеспечит удовлетворительных результатов локализации в случае изменения настроек [46], если только не может быть использован какой-либо соответствующий метод адаптации, что все еще остается открытой проблемой в DL в целом.

Метод, предложенный в [35], не показал способность обобщать и адаптироваться к расстоянию между центром решетки и источником. Для этого необходимо предложить новый метод классификации.

В [34] производительность предложенной модели локализации была изучена при условиях обучения, аналогичных сценарию целевого приложения. Однако способность модели локализации работать в помещении другого размера, чем в процессе обучения, не проверялась. С этой целью важно изучить возможность обобщения предлагаемой модели локализации, включив больше обучающих данных для помещений разного размера.

Большинство исследований в области локализации нескольких акустических источников были сосредоточены на решении этой проблемы как проблемы классификации, и было предложено множество моделей. Для проверки эффективности этих моделей в реальном времени пришлось значительно расширить набор данных (как пример, TAU-NIGENS Spatial Sound Events 2020) которые включают наиболее возможные случаи и сценарии, моделирующие реальную рабочую среду. Среди них отметим: статические и движущиеся реверберирующие акустические события, возможно перекрытие акустических событий во времени и пространстве, расположение акустических источников внутри одного или нескольких помещений с учетом импульсной характеристики помещения и его размеров, расположение микрофонной решетки внутри помещения, а также расстояние между акустическими источниками и центром микрофонной решетки, а также учет шума и реверберации. Большой размер набора данных, необходимый для метода глубокого обучения, требует больших вычислительных затрат и времени выполнения, которое может достигать десятков часов. Для решения этой проблемы возникла необходимость уменьшить размер набора данных и получить ту же производительность модели. Процесс уменьшения размера набора данных требует пренебрежения одним из ранее упо-

мянутых сценариев. С теоретической и практической точки зрения можно только уменьшить количество акустических источников, а это тот параметр, который больше всего влияет на размер набора данных.

На основании проведенного анализа предлагается разработать метод локализации акустического источника, способный обеспечить эффективность его работы при низких вычислительных затратах путем интеграции традиционных и современных методов для объединения преимуществ каждого из этих групп методов.

Чтобы построить эффективную модель локализации нескольких акустических источников в реальном времени с меньшими вычислительными затратами, предлагаем новый метод, который интегрирует разделение акустических источников с локализацией одного источника. Для построения этой модели изначально предлагаем применить эффективный метод разделения акустических источников, а затем разработанный метод локализации одного акустического источника с помощью двух ортогональных микрофонных решеток будет применен к каждому акустическому источнику из разделенных источников.

Выводы

Микрофонные решетки широко используются для локализации источников за счет получения пространственной информации с учетом временной задержки между микрофонами. В статье были представлены геометрические структуры этих решеток, в том числе линейная, плоская и сферическая. Несмотря на то что линейная микрофонная решетка проста по структуре и возможностям обработки, она ограничена неоднозначностью. Плоские решетки преодолевают неоднозначность «перед-зад», однако не справляются с неоднозначностью «верх-низ». Сферическая микрофонная решетка позволяет локализовать источники в любой точке пространства без пространственной неоднозначности.

Методы, используемые для локализации акустических источников, делятся на традиционные, основанные на обработке сигналов, и современные, основанные на глубоких нейронных сетях.

Несмотря на то что некоторые традиционные методы могут адаптироваться к наблюдаемым сигналам, все они зависят от принятых предположений и допущений о характере среды, о свойствах сигналов и т.д. Модели глубокого обучения явно не требуют ни одного из этих предположений, а вместо этого эффективно адаптируются к предоставленным обучающим данным. Однако это также является основным недостатком современных методов, поскольку они менее способны к обобщению и менее универсальны, чем традиционные методы.

Таким образом, можно сказать, что современные модели локализации, разработанные и обученные под конкретную конфигурацию и с невысокими вычислительными затратами, не обеспечат удовлетворительных результатов локализации в случае изменения настроек, если не будет использован какой-либо метод адаптации, что до сих пор остается открытой проблемой.

Дано обоснование необходимости развития новых методов локализации, а также интеграции традиционных и современных интеллектуальных методов локализации для объединения преимуществ каждого из этих групп методов.

Список литературы / References

[1] Сазонтов А.Г., Смирнов И.П. Локализация источника в акустическом волноводе с неточно известными параметрами с использованием согласованной обработки в модовом пространстве. *Акустический журнал*, 2019, 65(4), 540–555 [Sazontov A.G., Smirnov, I.P. Source localization in an acoustic waveguide with inaccurately known parameters using matched mode space processing, *J. Acoust.,* 65(4), 540–550 (in Rus.)]

[2] Mesaros A., Heittola T., Eronen A., Virtanen T. Acoustic event detection in real life recordings. *18th European Signal Processing Conference, IEEE.* 2010, 1267–1271.

[3] Reynolds D.A. Speaker identification and verification using Gaussian mixture speaker models, *Speech communication*, 1995, 17(1–2), 91–108.

[4] Aggarwal R., Singh J.K., Gupta V.K., Rathore S., Tiwari M., Khare A. Noise reduction of speech signal using wavelet transform with modified universal threshold, *International Journal of Computer Applications*, 2011, 20(5), 4–19.

[5] Шаход Д.М., Ибряева О.Л. Метод подавления акустического эха на основе рекуррентной нейронной сети и алгоритма кластеризации. *Вестник ЮУрГУ. Вычислительная математика и информатика*, 2022, 11(2), 43–58 [Shahoud Gh.M., Ibryaeva O.L. Method of an Acoustic Echo Suppression Based on Recurrent Neural Network and Clustering, *Bull. SUSU. Comput. Math. Soft. Eng.*, 2022, 11(2), 43–58 (in Rus.)]

[6] Sturdivant R.L., Chong E.K. Systems engineering baseline concept of a multispectral drone detection solution for airports, *IEEE Access*, 2017, 5, 7123–7138.

[7] Tehrani A.K.Z., Makkiabadi B., Parsayan A., Hozhabr S.H. Sound source localization using time differences of arrival; Euclidean distance matrices based approach. *2018 9th International Symposium on Telecommunications, IEEE*. 2017, 2017, 91–95.

[8] Chiariotti P., Martarelli M., Castellini P. Acoustic beamforming for noise source localization – Reviews, methodology and applications, *Mechanical Systems and Signal Processing*, 2019, 120, 422–448.

[9] Desai D., Mehendale N. A Review on Sound Source Localization Systems, *Archives of Computational Methods in Engineering*, 2022, 29(7), 4631–4642.

[10] Knapp C., Carter G. The generalized correlation method for estimation of time delay, *IEEE transactions on acoustics, speech, and signal processing*, 1976, 24(4), 320–327.

[11] Hirvonen T. Classification of spatial audio location and content using convolutional neural networks, *Audio Engineering Society Convention 138, Audio Engineering Society*, 2015.

[12] Nguyen T.N.T., Nguyen N.K., Phan H., Pham L., Ooi K., Jones D.L., Gan W.S. A general network architecture for sound event localization and detection using transfer learning and recurrent neural network. *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.* 2021, 935–939.

[13] Adavanne S., Politis A., Nikunen J., Virtanen T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, *IEEE Journal of Selected Topics in Signal Processing*, 2018, 13(1), 34–48.

[14] Yalta N., Nakadai K., Ogata T. Sound source localization using deep learning models, *Journal of Robotics and Mechatronics*, 2017, 29(1), 37–48.

[15] Nilanjan D., Amira S.A. *Direction of Arrival Estimation and Localization of Multi-Speech Sources*, Cham, Switzerland: Springer, 2018, 53.

[16] Lalan K. *Microphone Array Processing for Acoustic Source Localization in Spatial and Spherical Harmonics Domain*, *thesis submitted for the degree of doctor of philosophy*, *Indian Institute of Technology Kanpur*. Kanpur, 2015, 18.

[17] Siano D., Viscardi M., Panza M. A. Experimental acoustic measurements in far field and near field conditions: characterization of a beauty engine cover, *Recent Advances in Fluid Mechanics and Thermal Engineering*, 2014, 50–57.

[18] Cobos M., Antonacci F., Alexandridis A., Mouchtaris A., Lee B. A Survey of Sound Source Localization Methods in Wireless Acoustic Sensor Networks, *Wireless Communications and Mobile Computing*, 2017, 2017, 1–24.

[19] Sand S., Dammann A., Mensing C. *Positioning in wireless communications systems*, John Wiley & Sons, 2014, 280.

[20] Zhu N., Reza T. A modified cross-correlation algorithm to achieve the time difference of arrival in sound source localization, *Measurement and Control*, 2019, 52(3–4), 212–221.

[21] Rascon C., Meza I. Localization of sound sources in robotics: A review, *Robotics and Autonomous Systems*, 2017, 96, 184–210.

[22] Hosangadi R. A Proposed Method for Acoustic Source Localization in Search and Rescue Robot. *Proceedings of the 5th International Conference on Mechatronics and Robotics Engineering, ACM*. 2019, 134–140.

[23] Merino-Martínez R., Sijtsma P., Snellen M., Ahlefeldt T., Antoni J., Bahr C. J., Blacodon D., Ernst D., Finez A., Funke S., Geyer T. F. A review of acoustic imaging methods using phased microphone arrays: Part of the "Aircraft Noise Generation and Assessment" Special Issue, *CEAS Aeronautical Journal*, 2019, 10, 197–230.

[24] Schmidt R. Multiple emitter location and signal parameter estimation, *IEEE transactions on antennas and propagation*, 1986, 34(3), 276–280.

[25] Zhong Y., Xiang J., Chen X., Jiang Y., Pang J. Multiple Signal Classification-Based Impact Localization in Composite Structures Using Optimized Ensemble Empirical Mode Decomposition, *Applied Sciences*, 2018, 8(9), 1447.

[26] Фурлетов Ю. М. Классификация объектов и их действий методом анализа звуковых сигналов, *DSPA: ВОПРОСЫ ПРИМЕНЕНИЯ ЦИФРОВОЙ ОБРАБОТКИ СИГНАЛОВ Учредители: Российское научно-техническое общество радиотехники, электроники и связи им. АС Попова*, 2021, 11(4), 15–21 [Furletov Yu. M. Classification of objects and their actions by analyzing sound signals. DSPA: ISSUES OF APPLICATION OF DIGITAL SIGNAL PROCESSING Founders: Russian Scientific and Technical Society of Radio Engineering, Electronics and Communications named. AS Popova, 2021, 11(4), 15–21 (in Rus.)]

[27] Grumiaux P. A., Kitić S., Girin L., Guérin A. A survey of sound source localization with deep learning methods, *The Journal of the Acoustical Society of America*, *152*(1), 107–151.

[28] Chakrabarty S., Habets E. A. Multi-speaker DOA estimation using deep convolutional networks trained with noise signals, *IEEE Journal of Selected Topics in Signal Processing*, 2019, 13(1), 8–21.

[29] He W., Motlicek P., Odobez J. M. Deep neural networks for multiple speaker detection and localization. *2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE*. Brisbane, Australia, 2018, 74–79.

[30] Thuillier E., Gamper H., Tashev I. J. Spatial audio feature discovery with convolutional neural networks. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE*. Calgary, Canada, 2018, 6797–6801.

[31] Vera-Diaz J.M., Pizarro D., Macias-Guarasa J. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates, *Sensor*, 2018, 18(10), 3418.

[32] Nguyen T. N.T., Gan W. S., Ranjan R., Jones D. L. Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, 28, 2626–2637.

[33] Chakrabarty S., Habets E. A. Multi-scale aggregation of phase information for complexity reduction of CNN based DOA estimation. *2019 27th European Signal Processing Conference (EUSIPCO), IEEE*. 2019, 1–5.

[34] Liu N., Chen H., Songgong K., Li Y. Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays, *J. Acoust. Soc. Am*., 2021, 1069–1084.

[35] Giovanni A., Roberto A. *Investigating the generalization abilities of a deep learning method for sound source localization using small-sized microphone arrays, Project Course – M. Sci. on Music and Acoustic Engineering, Politecnico di Milano*. Italy, 2022.

[36] Lin Y., Wang Z. A report on sound event localization and detection, *Detection Classification Acoust. Scenes Events Challenge, Tech. Rep.*, 2019.

[37] Lu Z. Sound event detection and localization based on CNN and LSTM, *Detection Classification Acoust. Scenes Events Challenge, Tech. Rep.*, 2019.

[38] Maruri H. C., Meyer P. L., Huang J., Ontiveros J. A.D.H., Lu H. Gcc-phat cross-correlation audio features for simultaneous sound event localization and detection (seld) on multiple rooms, *DCASE 2019 Challenge, Tech. Rep.*, 2019.

[39] Zhang J., Ding W., He L. Data augmentation and prior knowledge-based regularization for sound event localization and detection, *DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge*, 2019.

[40] Xue W., Ying T., Chao Z., Guohong D. Multi-beam and multi-task learning for joint sound event detection and localization, *DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge*, 2019.

[41] Cao Y., Kong Q., Iqbal T., An F., Wang W., Plumbley M. D. Polyphonic sound event detection and localization using a two-stage strategy, *arXiv preprint arXiv:1905.00268,* 2019.

[42] Tian C. Multiple CRNN for SELD, *parameters*, 2020, 488211(508257), 490326.

[43] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, 770–778.

[44] 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*. 2019, 770–774.

[45] Suvorov D., Dong G., Zhukov R. Deep residual network for sound source localization in the time domain, *arXiv preprint arXiv:1808.06429*, 2018.

[46] Le Moing G., Vinayavekhin P., Agravante D. J., Inoue T., Vongkulbhisal J., Munawar A., Tachibana R. Data-efficient framework for real-world multiple sound source 2D localization. *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*. 2021, 3425–3429.