DOI: 10.62994/DVO.2025.95.10.077

ОБНАРУЖЕНИЕ ОДИНОЧНОГО ОБЪЕКТА В РЕЖИМЕ ШУМОПЕЛЕНГОВАНИЯ ПРИ ПОМОЩИ НЕЙРОСЕТЕВОГО МЕТОДА ДИСТИЛЛЯЦИИ ЗНАНИЙ

Кочегаров Н.К.[1], **Иванов М.П.**[2], **Шпак Ю.В.**[1]

[1]*Дальневосточный федеральный университет, г. Владивосток*
[2]*Тихоокеанский океанологический институт им. В.И. Ильичева, г. Владивосток*
kochegarov_nk@dvfu.ru

В данной работе предложен теоретический подход к обнаружению и пеленгованию объекта по шумовым акустическим сигналам, возникающим при его перемещении в замкнутом объёме (бухты, заливы). Для упрощения моделирования и подготовки набора данных морская вода была замена на воздух, замкнутый объём морской среды на помещение размером 5x7 метров, в качестве одиночного объекта выбран человек. Для реализации задачи используется метод дистилляции знаний между двумя нейронными сетями: обучаемой аудиосетью (ученик) и уже обученной зрительной сетью (учитель) [1]. Предполагается, что система работает в условиях ограниченной видимости или с ограничениями на использование видеонаблюдения, поэтому основным источником информации являются микрофоны. Зрительная сеть, принимающая изображения с камеры, используется исключительно для формирования эталонных меток во время обучения аудиосети. Описаны архитектура сети, процесс синхронизации данных, функция потерь и логика корректировки параметров аудиосети. Подробно раскрыта математическая модель дистилляции знаний, включающая раздельную оценку вероятности обнаружения и угла пеленга [2].

Предлагаемая система состоит из двух компонентов:

- **Аудиосеть (ученик)** — нейронная сеть, принимающая на вход многоканальные аудиосигналы от массива микрофонов и выводящая:
    1. Вероятность наличия человека (0 или 1)
    2. Угол пеленга на человека в градусах (например, от 0° до 360°)

- **Зрительная сеть (учитель)** – предобученная свёрточная нейронная сеть, принимающая на вход кадры с видеокамеры и выводящая:
  - Вероятность наличия человека
  - Позицию человека в поле зрения камеры, которая может быть переведена в угол пеленга относительно точки установки системы

Обе сети работают параллельно и синхронизированы по времени.

Для формализации процесса дистилляции знаний определим формальные переменные и операции, используемые в системе.

Пусть:

$A_t = (p_a, \theta_a)$ – выход аудиосети в момент времени t

$V_t = (p_v, \theta_v)$ – выход зрительной сети в момент времени t,

где:

$p_a \in [0, 1]$ – вероятность того, что человек обнаружен аудиосетью

$p_v \in [0, 1]$ – вероятность того, что человек обнаружен зрительной сетью

$\theta_a \in [0°, 360°]$ – оценка пеленга, выданная аудиосетью

$\theta_v \in [0°, 360°]$ – оценка пеленга, выданная зрительной сетью.

Бинарный результат обнаружения формируется на основе порога $\tau \in (0, 1)$:

$$\hat{p}_a = \begin{cases} 1, & p_a > \tau \\ 0, & p_a < \tau \end{cases}$$

Аналогично для $\hat{p}_v$.

Обучение аудиосети происходит только в случае расхождения между $\hat{p}_a$ и $\hat{p}_v$, т.е. если $\hat{p}_a \neq \hat{p}_v$.

Такие случаи могут быть двух типов:

1. Ложно-отрицательный: $\hat{p}_a = 0, \quad \hat{p}_v = 1$

2. Ложно-положительный: $\hat{p}_a = 1, \quad \hat{p}_v = 0$

Однако второй случай встречается значительно реже, поскольку зрительная сеть заранее обучена и более надёжна [3]. Поэтому в большинстве ситуаций мы рассматриваем первый случай, когда аудиосеть не обнаруживает человека, который реально присутствует.

Функция потерь определяется следующим образом:

$$L_t \begin{cases} 0, & \hat{p}_a = \hat{p}_v \\ \alpha \cdot L_{cls}(p_a, p_v) + \beta \cdot L_{dir}(\theta_a, \theta_v), & \hat{p}_a \neq \hat{p}_v \end{cases}$$

Где:

- L_{cls} – бинарная кросс-энтропия для классификации наличия человека:

$$L_{cls}(p_a, p_v) = -[\, p_v \, log(p_a) + (1 - p_v) \, log(1 - p_a)]$$

- L_{dir} – угловая потеря, например, среднеквадратичная ошибка (MSE):

$$L_{dir}(\theta_a, \theta_v) = (\theta_a - \theta_v)^2$$

Коэффициенты α и β позволяют задать относительную важность классификационной и пеленгационной части. Их можно подбирать эмпирически или использовать автоматическую настройку в зависимости от динамики обучения [3].

На каждом шаге *t*, где выполняется условие $\hat{p}_a \neq \hat{p}_v$, аудиосеть обновляет свои параметры ω_a по правилу стохастического градиентного спуска:

$$\omega_a^{(t+1)} = \omega_a^{(t)} - \eta \Delta \omega_a L_t$$

Где:

- η — скорость обучения (learning rate)
- $\nabla \omega_a L_t$ — градиент функции потерь по параметрам аудиосети

Скорость обучения может быть постоянной или изменяться динамически, например, по стратегии убывающего learning rate или циклической схеме.

Для обработки звуковых сигналов предлагается использовать трёхмерную свёрточную нейросеть, работающую с временно-частотным представлением сигнала (например, STFT). Входом служат спектрограммы нескольких микрофонов, объединённые в 3D-тензор. Выход состоит из двух составляющих: одна для классификации наличия человека, другая — для регрессии угла пеленга.

В качестве зрительной сети предлагается использовать предобученную модель YOLOv8, дополненную модулем оценки направления на объект [2]. Камера устанавливается таким образом, чтобы охватывала тот же сектор, что и микрофонный массив. Пе-

ленг рассчитывается как горизонтальный угол между центральным лучом камеры и положением человека в кадре.

Обучение аудиосети предполагается осуществлять в реальном времени, с использованием mini-batch'ей, собранных за последние несколько секунд. Оптимизатор: AdamW, начальная скорость обучения: 1e-4. Также предполагается применять стратегию снижения скорости обучения при стабилизации потерь.

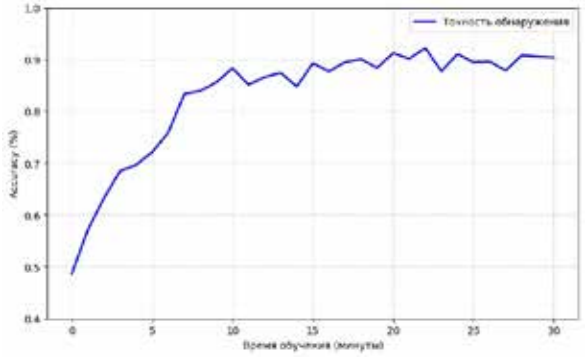

Рис. 1. График точности обнаружения от времени обучения

Моделирование показало, что даже при высоком уровне шума и слабых акустических сигналах, аудиосеть способна достигнуть высокой точности обнаружения и пеленгования.

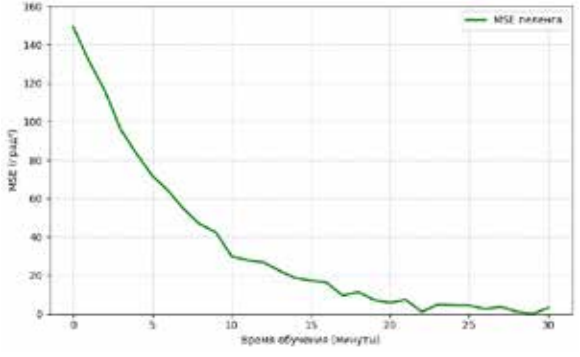

Рис. 2. График среднеквадратичной ошибки пеленга от времени обучения

Точность обнаружения резко возрастает в первые минуты обучения, особенно интенсивно — в течение первых 10–15 минут. После этого график выходит на плато, стабилизируясь на уровне около 95%. Это свидетельствует о том, что аудиосеть успешно адаптируется к характеристикам звука, создаваемого человеком, и к окружающей шумовой обстановке.

Предложенный метод дистилляции знаний позволяет обучить аудиосеть без явной разметки, используя визуальные данные от камеры как «идеальный» эталон. Такой подход открывает возможность создания систем обнаружения и пеленгования, работающих только на основе звука, что особенно актуально в условиях ограниченной видимости.

Основные преимущества метода:

- Отказ от ручной разметки данных.
- Возможность обучения в реальном времени.
- Гибкость к изменению условий окружающей среды.

Ограничения:

- Необходимость наличия камеры на этапе обучения.
- Сложность синхронизации аудио- и видеосигналов.
- Ограниченная применимость в случаях, когда человек находится вне поля зрения камеры.

Литература

1. Hinton G., Vinyals O., Dean J. Distilling the knowledge in a neural network // arxiv.org. 2015. DOI: 10.48550/arXiv.1503.02531.
2. Redmon J., Farhadi A. YOLOv3: An incremental improvement // arxiv.org. 2018. DOI: 10.48550/arXiv.1804.02767.
3. Hershey S., Chaudhuri S., Ellis D.P.W., Gemmeke J.F., Jansen A., Moore R.C., Plakal M., Platt D., Saurous R.A., Seybold B., Slaney M., Weiss R.J., Wilson K. CNN architectures for large-scale audio classification // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. P. 131-135.