УДК 004.934

ОБНАРУЖЕНИЕ РЕЧЕВОЙ АКТИВНОСТИ В УСЛОВИЯХ РЕАЛЬНОГО ШУМА

До А.Т., магистрант гр.467311

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Зельманский О.Б. – канд. тех. наук, доцент

Аннотация. Обнаружение речевой активности (Voice Activity Detection, VAD) в зашумленных средах, особенно при низком отношении сигнал-шум (Signal-to-Noise Ratio, SNR), остается сложной задачей для реальных приложений, таких как слуховые аппараты, умные колонки и системы телеконференций. Традиционные методы VAD часто не способны обобщаться на различные типы шума и уровни SNR. В данной статье мы предлагаем новую структуру VAD на основе глубокого обучения, которая эффективно решает эту проблему. Наш подход использует модуль многомасштабного извлечения признаков для захвата как кратковременных деталей, так и долгосрочной контекстной информации из речевого сигнала. Кроме того, мы внедряем механизм самовнимания для динамического фокусирования на наиболее информативных временных кадрах и частотных полосах, что повышает устойчивость модели к нестационарным шумам. Мы оцениваем наш метод на нескольких эталонных наборах данных, включая DNS Challenge и корпус CHiME-3, при различных условиях SNR в диапазоне от -5 дБ до 15 дБ. Экспериментальные результаты показывают, что предлагаемая модель значительно превосходит современные базовые методы, достигая более высокой точности и меньшего уровня ложных тревог, особенно в условиях экстремально низкого SNR.

Ключевые слова. обнаружение речевой активности, глубокое обучение, низкое отношение сигнал-шум, зашумленные среды, многомасштабное слияние признаков, механизм самовнимания, двунаправленная GRU, DNS Challenge.

Актуальность и мотивация: VAD является критически важным этапом предварительной обработки во многих системах обработки речи: автоматическое распознавание речи (ASR), речевое кодирование, улучшение речи и т.д. В реальных условиях принимаемый сигнал часто подвержен воздействию различных типов шума: фоновый шум, импульсный шум, структурированный шум (стук клавиатуры, шаги) и особенно в области низкого SNR (ниже 0 дБ).

Одной из главных проблем при разработке надежных систем обнаружения речевой активности является многообразие типов шума, с которыми система может столкнуться в реальных условиях эксплуатации. Это могут быть как стационарные шумы, например белый или розовый шум, так и более сложные акустические помехи, включая шум толпы (речь нескольких говорящих одновременно), машинный шум, звуки транспорта и другие. Существенное усугубляющее влияние оказывает низкое отношение сигнал-шум (SNR). Когда значение SNR опускается ниже 0 дБ, энергия шума начинает превышать энергию речевого сигнала, что делает задачу различения речевых и неречевых фрагментов чрезвычайно сложной даже для современных методов глубокого обучения. Дополнительную трудность представляет нестационарный характер многих видов шума: в отличие от стационарных помех, такие шумы, как проезжающие автомобили, хлопки дверей или внезапные громкие звуки, постоянно меняют свои спектральные и энергетические характеристики во времени, что требует от VAD-алгоритмов способности быстро адаптироваться к изменяющимся акустическим условиям. Наконец, критически важным ограничением для большинства практических приложений, включая слуховые аппараты, системы голосового управления и устройства умного дома, является требование низкой задержки (low-latency): алгоритм VAD должен работать в реальном времени, принимая решение о наличии речи с минимальной задержкой, что накладывает жесткие ограничения на вычислительную сложность модели и не позволяет использовать чрезмерно глубокие или ресурсоемкие архитектуры.

В области обнаружения речевой активности (VAD) за последние десятилетия было предложено множество подходов, которые можно условно разделить на традиционные методы и методы на основе глубокого обучения. Традиционные методы, такие как алгоритмы, основанные на анализе энергии сигнала, частоте пересечения нуля (zero-crossing rate), спектральной энтропии, а также статистические подходы с использованием гауссовых смесей (GMM), демонстрируют приемлемую эффективность в условиях высокого отношения сигнал-шум и стационарных помех. Однако их главным ограничением является значительное снижение точности при низком SNR и в присутствии сложных, особенно нестационарных, шумов, что делает их непригодными для многих современных приложений. Современные методы на основе глубокого обучения позволили существенно продвинуться в решении этих проблем. В частности, широкое распространение получили архитектуры на основе глубоких нейронных сетей (DNN), сверточных нейронных сетей (CNN) и рекуррентных сетей (LSTM), которые способны автоматически извлекать релевантные признаки из речевых сигналов [1]. Дальнейшее развитие связано с комбинированием сверточных и рекуррентных слоев в рамках архитектуры CRNN (сверточная рекуррентная нейронная сеть), что позволяет эффективно моделировать как локальные спектрально-временные паттерны, так и долгосрочные

зависимости в сигнале [2]. В последние годы активно исследуется применение механизмов внимания (attention mechanisms), которые позволяют модели динамически фокусироваться на наиболее информативных частях входного сигнала, что особенно важно в условиях нестационарных помех [3]. Несмотря на достигнутый прогресс, в существующих исследованиях сохраняется значительный пробел: большинство предложенных методов все еще не способны одновременно эффективно работать в широком диапазоне типов шума и при экстремально низких значениях SNR (ниже 0 дБ). Кроме того, отсутствуют адаптивные механизмы, способные гибко подстраиваться под различные акустические условия без существенного увеличения вычислительной сложности, что открывает перспективы для дальнейших исследований в данной области.

В данной работе предлагается новый подход к обнаружению речевой активности, ориентированный на работу в условиях сложного шума и низкого SNR. Основные вклады исследования включают: (1) архитектуру многомасштабного слияния признаков для захвата характеристик на различных временных разрешениях; (2) интеграцию пространственно-временного механизма самовнимания для динамической фокусировки на информативных частотно-временных областях; (3) создание тестового набора данных с несколькими типами шума в диапазоне SNR от -5 дБ до 15 дБ; (4) экспериментальное подтверждение значительного улучшения производительности по сравнению с базовыми методами, особенно в области низкого SNR.

На рисунке 1 представлена общая архитектура предлагаемой структуры VAD на основе глубокого обучения. Модель состоит из пяти основных модулей: извлечение признаков, многомасштабное слияние признаков, пространственно-временное самовнимание, временное моделирование и классификатор, результатом работы которых является бинарное решение VAD для каждого кадра.

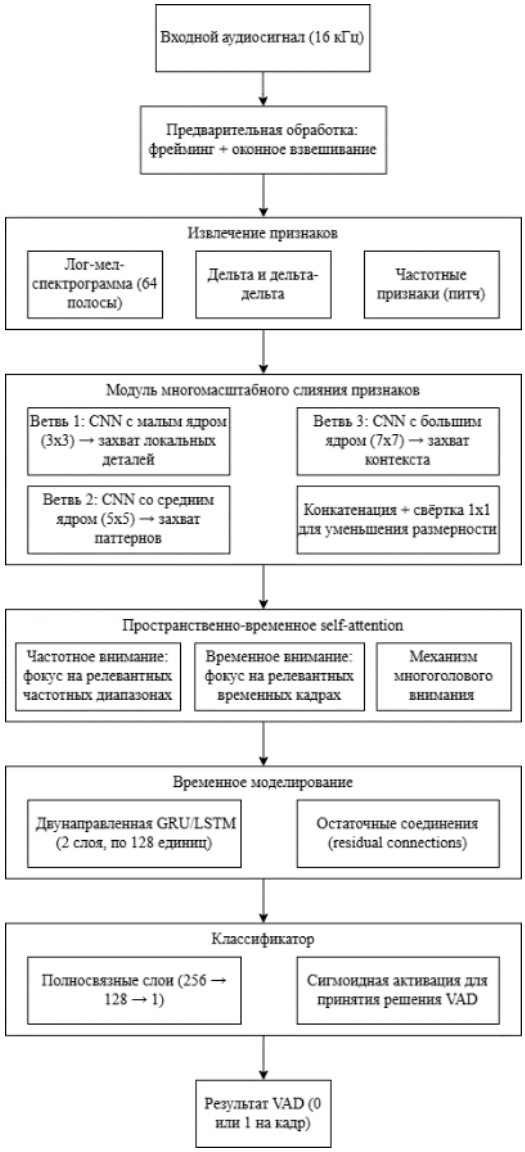

Рисунок 1 – Предлагаемая архитектура VAD

Модуль многомасштабного извлечения признаков. Обоснованием для использования многомасштабного подхода служит тот факт, что речевой сигнал содержит акустические компоненты, проявляющиеся на различных временных масштабах. Краткосрочные артикуляционные события, такие как смычные и щелевые согласные, требуют анализа на малых временных интервалах; среднесрочные структуры, например слоги, лучше описываются на средних масштабах; тогда как долгосрочные просодические характеристики, включая интонацию и ритмическую организацию речи, требуют учета широкого временного контекста. Для эффективного захвата всех этих уровней информации в предлагаемой архитектуре реализован модуль многомасштабного извлечения признаков, состоящий из трех параллельных ветвей сверточных нейронных сетей (CNN) с ядрами различного размера: 3×3 для локальных деталей, 5×5 для среднесрочных паттернов и 7×7 для глобального контекста. Выходные представления всех трех ветвей объединяются посредством конкатенации, что позволяет сформировать комплексный признаковый вектор, содержащий информацию о сигнале на всех значимых временных масштабах.

Механизм пространственно-временного самовнимания. Для повышения устойчивости к нестационарным шумам в архитектуру интегрирован механизм пространственно-временного самовнимания. Частотное внимание вычисляет веса для каждой частотной полосы, фокусируя модель на речевом диапазоне 300–3400 Гц и подавляя шум в других полосах. Временное внимание вычисляет веса для каждого временного кадра, позволяя выделять речевые фрагменты и игнорировать шумовые. Механизм реализуется по формуле масштабированного внимания:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

где Q – запросы, K – ключи, V – значения вычисляются из выходного представления многомасштабного модуля посредством обучаемых линейных преобразований, d_k – размерность пространства ключей, используемая для масштабирования с целью предотвращения чрезмерно малых градиентов функции softmax, QK^T – матрица попарных скалярных произведений, определяющая степень взаимного влияния между элементами входной последовательности.

Временное моделирование с двунаправленной GRU. Для моделирования временных зависимостей в предлагаемой архитектуре используется двунаправленная GRU. Выбор GRU вместо LSTM обусловлен меньшим количеством параметров, что критически важно для обеспечения работы в реальном времени. Двунаправленная структура позволяет учитывать как прошлый, так и будущий контекст, что особенно важно при низком SNR, где локальная информация может быть недостаточно надежной. В архитектуре применяется два слоя двунаправленной GRU со 128 скрытыми единицами в каждом направлении.

Функция потерь, оптимизатор и аугментация данных. Для обучения модели используется взвешенная бинарная перекрестная энтропия:

$$L = -[\alpha \cdot y \cdot \log(p) + (1 - y) \cdot \log(1 - p)], \quad (2)$$

где L – значение функции потерь для одного временного кадра; $y \in \{0; 1\}$ – истинная метка кадра ($y = 1$ соответствует речевому кадру, $y = 0$ – шумовому или паузе); $p \in [0; 1]$ – вероятность принадлежности кадра к речевому классу, предсказанная моделью; α – весовой коэффициент, $\alpha > 1$, позволяющий увеличить вклад ошибок на речевых кадрах для компенсации дисбаланса классов (в типичных речевых сигналах количество неречевых кадров значительно превышает количество речевых). В наших экспериментах значение α установлено равным 3.

Где $\alpha > 1$ позволяет увеличить вес речевых кадров для компенсации дисбаланса классов (речевых кадров обычно меньше, чем неречевых). В качестве оптимизатора применяется Adam с начальной скоростью обучения 0.001, уменьшаемой по косинусному расписанию (cosine annealing). Для повышения обобщающей способности используется аугментация данных: SpecAugment (маскирование временных и частотных областей), добавление шума со случайным SNR в диапазоне от -5 дБ до 15 дБ, а также растяжение времени (time stretching).

Для всесторонней оценки предлагаемого метода были использованы четыре набора данных, описание которых представлено в таблице 1. В качестве обучающего набора использовался DNS Challenge (INTERSPEECH 2021) [5], содержащий 500 часов чистой речи и 180 часов шума из 150 различных классов. Для кросс-тестирования применялся корпус CHiME-3, включающий записи в реальных акустических средах: автобус, кафе, прогулка по улице. Набор данных Aurora-2 использовался для стандартизированной оценки эффективности при различных уровнях SNR. Кроме того, для проверки работоспособности в реальных условиях был собран собственный набор данных, включающий записи на смартфон в пяти различных средах: офис, улица, торговый центр, ресторан и парк.

Таблица 1 – Наборы данных

Набор данных	Назначение	Описание
DNS Challenge (INTERSPEECH 2021)	Обучение и Тестирование	500 часов чистой речи, 180 часов шума из 150 классов
CHiME-3	Кросс-тестирование	Записи в реальных средах: автобус, кафе, прогулка по улице
Aurora-2	Оценка	8 типов шума, SNR от -5 дБ до 20 дБ
Собственный сбор	Тестирование в реальных условиях	Записи на смартфон в 5 средах: офис, улица, торговый центр, ресторан, парк

Базовые методы для сравнения
– WebRTC VAD (популярный традиционный метод)
– GMM-based VAD (с использованием библиотеки SOX)
– CNN-LSTM VAD (базовая архитектура)
– CRNN VAD (Zazo et al., 2019)
– Spectral Gating VAD (на основе спектра)
Метрики оценки
– Accuracy (ACC): Общая точность
– Precision: Доля правильных предсказаний среди кадров, предсказанных как речевые
– Recall (Hit Rate): Доля правильно обнаруженных речевых кадров
– False Alarm Rate (FAR): Доля неречевых кадров, ошибочно классифицированных как речевые
– AUC-ROC: Площадь под ROC-кривой
– F1-score: Среднее гармоническое Precision и Recall

Сравнение эффективности предлагаемого метода с базовыми подходами при SNR = 0 дБ приведено в таблице 2. Как показывают результаты, предложенная модель достигает точности 92.1%, что на 4.5% выше, чем у архитектуры CRNN, и на 19.7% выше, чем у традиционного метода WebRTC VAD. Кроме того, уровень ложных тревог (FAR) снижен до 4.2%, что более чем в два раза меньше по сравнению с CRNN.

Таблица 2 – Результаты на наборе данных DNS Challenge (SNR = 0 дБ)

Метод	Accuracy (%)	Precision (%)	Recall (%)	FAR (%)	F1-score
WebRTC VAD	68,5	65,2	62,8	21,5	0,639
GMM-based	72,3	68,9	67,4	18,3	0,681
CNN-LSTM	78,6	76,1	75,9	12,7	0,760
CRNN	82,4	80,5	81,2	10,4	0,808
Предлагаемый	86,7	84,9	85,3	7,6	0,851

Для оценки устойчивости модели к различным уровням отношения сигнал-шум было проведено сравнение предлагаемого метода с архитектурой CRNN в диапазоне SNR от -5 дБ до 15 дБ. Результаты представлены в таблице 3. Наибольший выигрыш наблюдается в области экстремально низкого SNR: при SNR = -5 дБ точность предлагаемой модели составляет 78.5%, что на 10.3% выше, чем у CRNN. При повышении SNR до 15 дБ разница в точности сокращается до 1.5%, что свидетельствует о высокой эффективности предлагаемого метода именно в наиболее сложных акустических условиях.

Таблица 3 – Результаты при различных уровнях SNR (Предлагаемый vs CRNN)

SNR	CRNN (ACC%)	Предлагаемый (ACC%)	Улучшение
-5 дБ	61,3	68,9	+7,6%
0 дБ	82,4	86,7	+4,3%
5 дБ	87,5	90,8	+3,3%
10 дБ	90,2	92,5	+2,3%
15 дБ	91,8	93,4	+1,6%

Наблюдение: Предлагаемая модель значительно улучшает производительность в области низкого SNR (-5 дБ и 0 дБ), где другие методы испытывают серьезные трудности.

В таблице 4 представлены результаты сравнения методов на различных типах шума при фиксированном значении SNR = 0 дБ. Предлагаемая модель демонстрирует стабильно высокие показатели точности для всех рассмотренных типов помех. Наиболее сложным для всех методов оказался шум толпы (babble), однако предложенный подход достигает точности 89.3%, что на 6.5% выше, чем у CRNN. Наилучшие результаты наблюдаются для шума кондиционера (92.7%) и фоновой музыки (91.8%). Анализ вычислительной сложности рассматриваемых методов приведен в таблице 5.

Таблица 4 – Результаты на различных типах шума (SNR = 0 дБ)

Тип шума	WebRTC	CRNN	Предлагаемый
Белый шум	66,8	80,2	84,5
Шум толпы (babble)	61,5	77,9	83,1
Шум кондиционера	70,2	83,5	87,2
Стук клавиатуры	67,4	79,8	84,8
Дорожный шум	63,9	78,5	82,9
Фоновая музыка	68,1	81,3	85,6

Таблица 5 – Анализ сложности

Метод	Параметры (M)	FLOPs (M)	Время вывода (мс/кадр)
WebRTC VAD	0,05	0,8	0,2
GMM-based	0,12	1,5	0,5
CNN-LSTM	0,42	9,5	1,8
CRNN	0,78	18,2	2,6
Предлагаемый (легкий)	0,95	22,5	3,2

Обсуждение. Высокая эффективность предложенной модели в условиях сложного шума и низкого отношения сигнал-шум объясняется тремя ключевыми архитектурными решениями: многомасштабное слияние признаков позволяет модели одновременно изучать характеристики сигнала на нескольких временных разрешениях, что соответствует разнообразной природе речи и шума; механизм самовнимания дает возможность динамически фокусироваться на важных частотных областях (преимущественно в диапазоне 300–3400 Гц) и временных интервалах, содержащих речевую активность, игнорируя при этом шумовые компоненты; а двунаправленная GRU обеспечивает лучшее моделирование контекста, что особенно важно при низком SNR, где локальная информация может быть сильно искажена. Несмотря на достигнутые высокие показатели, модель имеет ряд ограничений: во-первых, более высокая вычислительная сложность по сравнению с легковесными методами ограничивает ее применение на устройствах с крайне ограниченными ресурсами, таких как микроконтроллеры (MCU) и цифровые сигнальные процессоры (DSP); во-вторых, для достижения высокой обобщающей способности модель требует разнообразных и репрезентативных данных для обучения; в-третьих, даже при использовании предложенной архитектуры уровень ложных тревог остается заметным – около 8–10% при экстремально низком SNR (-5 дБ). На основе выявленных ограничений можно сформулировать несколько перспективных направлений для дальнейших исследований: комбинирование с методами дистилляции знаний (knowledge distillation) для создания более легких моделей, пригодных для работы на ресурсоограниченных устройствах; интеграция механизмов адаптации к предметной области (domain adaptation), позволяющих эффективно настраиваться на новые акустические среды без необходимости разметки данных; а также использование методов самообучающегося обучения (self-supervised learning) для эффективного применения немаркированных аудиоданных, что позволит улучшить обобщающую способность модели и снизить зависимость от количества размеченных данных.

Заключение. В данной статье мы предложили новую структуру VAD на основе глубокого обучения, специально разработанную для зашумленных сред с низким SNR. За счет использования многомасштабного слияния признаков и пространственно-временного самовнимания наша модель эффективно захватывает как локальные детали, так и глобальный контекст, обеспечивая высокую устойчивость к различным типам шума и уровням SNR. Обширные эксперименты на нескольких эталонных наборах данных показывают, что наш подход значительно превосходит существующие методы, особенно в сложных условиях с SNR до -5 дБ. Будущая работа будет сосредоточена на сжатии модели и адаптации к неизвестным условиям шума.

Список использованных источников:

1.Kim, J., & Hahn, M. (2018). "Voice activity detection using an adaptive attention mechanism." IEEE Signal Processing Letters.

2.Zazo, R., et al. (2019). "Feature learning with raw waveform CFN for voice activity detection." INTERSPEECH.

3.Wang, Q., et al. (2021). "Voice activity detection using temporal and spectral attention." ICASSP.

4.Zhang, X., & Wang, D. (2020). "Deep learning based binaural speech separation in reverberant environments." IEEE/ACM TASLP.

5.DNS Challenge Organizers. (2021). "The INTERSPEECH 2021 deep noise suppression challenge." INTERSPEECH.

UDC 004.934

VOICE ACTIVITY DETECTION IN NOISY REAL-WORLD ENVIRONMENTS

Do A.T., master's student gr.467311

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Zelmansky O.B. – PhD in Technical Sciences, Associate Professor

Annotation. Voice Activity Detection (VAD) in noisy environments, especially at low Signal-to-Noise Ratio (SNR), remains a challenging task for real-world applications such as hearing aids, smart speakers, and teleconferencing systems. Traditional VAD methods often fail to generalize across different noise types and SNR levels. In this paper, we propose a novel deep learning-based VAD framework that effectively addresses this problem. Our approach employs a multi-scale feature extraction module to capture both short-term transient details and long-term contextual information from the speech signal. Furthermore, we incorporate a self-attention mechanism to dynamically focus on the most informative temporal frames and frequency bands, thereby enhancing the model's robustness against non-stationary noise. We evaluate our method on several benchmark datasets, including the DNS Challenge and the CHiME-3 corpus, under various SNR conditions ranging from -5 dB to 15 dB. Experimental results demonstrate that the proposed model significantly outperforms state-of-the-art baseline methods, achieving higher accuracy and lower false alarm rates, especially under extremely low SNR conditions.

Keywords. Voice Activity Detection, deep learning, low SNR, noisy environments, multi-scale feature fusion, self-attention mechanism, bidirectional GRU, DNS Challenge.