С

УДК 004.8:004.6

И. Н. Савенков, Т. В. Ермоленко, А. В. Цыбик
Государственное образовательное учреждение высшего профессионального образования
«Донецкий национальный университет», г. Донецк
83001, г. Донецк, ул. Университетская, 24

РАЗРАБОТКА VAD-АЛГОРИТМА НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ

I. N. Savenkov, T. V. Yermolenko, A. V. Tsybik
State Educational Institution of Higher Professional Education «Donetsk National University»
83001, Donetsk, University st, 24

DEVELOPING A VAD-ALGORITHM BASED ON DEEP LEARNING

І. М. Савенков, Т. В. Єрмоленко, О. В. Цибік
Державна освітня установа вищої професійної освіти
«Донецький національний університет», м. Донецьк
83001, м. Донецьк, вул. Університетська, 24

РОЗРОБКА VAD-АЛГОРИТМА НА БАЗІ ГЛИБОКОГО НАВЧАННЯ

В статье дан обзор признаков, по которым определяется наличие речевой составляющей в аудиосигнале, а также наиболее известных алгоритмов, детектирующих речь. Для классификации фреймов сигнала на классы «шум»/«речь» предложена архитектура сверточной сети, на вход которой поступает изображение спектрограммы фрейма. Обучение и тестирование сети проводилось на наборе данных с разными видами шумовых эффектов, взятых с корпусов аудиоданных, находящихся в свободном доступе.
Ключевые слова: детектор речевой активности, акустический шум, энергия спектра, спектрограмма, сверточная нейронная сеть.

The article provides an overview of the signs by which the presence of a speech component in the audio signal is determined, as well as the most well-known algorithms that detect speech. To classify signal frames into "noise"/"speech" classes, a convolutional network architecture is proposed, the input of which receives an image of the spectrogram of the frame. Training and testing of the network was carried out on a data set with different types of noise effects taken from audio data bodies that are freely available.
Key words: Voice Activity Detector, acoustic noise, spectrum energy, spectrogram, convolutional neural network.

У статті дано огляд ознак, за якими визначається наявність мовленєвої складової в аудіосигналі, а також найбільш відомих алгоритмів, що детектують мову. Для класифікації фреймів сигналу на класи «шум»/»мова» запропонована архітектура згорткової мережі, на вхід якої надходить зображення спектрограми фрейму. Навчання і тестування мережі проводилося на наборі даних з різними видами шумових ефектів, взятих з корпусів аудіоданих, що знаходяться у вільному доступі.
Ключові слова: детектор мовленнєвої активності, акустичний шум, енергія спектра, спектрограма, згорткова нейронна мережа.

Введение

В последнее время произошел настоящий прорыв в области речевых технологий, активно вышедших на массовый рынок коммерческих продуктов. Основная причина этого связана с использованием методов глубокого обучения и наличием в открытом доступе размеченных аудиокорпусов для обучения нейросетей, распознающих речь человека. В системах, использующих речевые технологии, детектор голосовой активности (*Voice Activity Detector, VAD*) играет важную роль, позволяя обнаружить присутствие речи в аудиосигнале. Информация, полученная VAD, может быть использована для разделения аудиоинформации на голосовую часть и фоновый шум, что делает VAD неотъемлемой частью таких устройств как профессиональные диктофоны, системы преобразования речи в текст, системы голосового управления, системы связи, системы обработки речевого сигнала в слуховых аппаратах, усиливающие желаемый речевой сигнал и подавляющие шумовые компоненты.

Несмотря на обилие предложенных VAD-алгоритмов, они сталкиваются с общими проблемами, связанными со сложностью структуры речевого сигнала, в результате чего его невозможно описывать с помощью математических моделей, а также разнообразием внешних факторов, влияющих на запись и передачу голоса: VAD в большинстве случаев необходимо обнаруживать присутствие речи в искаженном шумом звуковом сигнале.

К факторам, усложняющим качественное выделение границ речи в звуковом сигнале, можно отнести присутствующие в фонограммах акустические шумы [1], плавное нарастание мощности речевого сигнала в начале речевой активности, начало речевой активности с шумного глухого щелевого звука (*с*, *ш*, *ф*, *х*, *ч*, *ц*) или шумного смычного звука (*т*, *к*, *п*, *б*, *д*, *г*), интенсивность и/или длительность которого невелика (не превышает 20 – 25 мс).

Указанные проблемы определяют основные недостатки существующих VAD-алгоритмов: недостаточная точность определения границ речевых сегментов и значительное ухудшение работы при наличии шумов.

Исходя из вышеизложенного, можно сделать вывод об актуальности повышения точности VAD-алгоритмов.

В данной работе проведен обзор признаков для выделения речи в сигнале, алгоритмов, детектирующих речевую активность, а также предложена нейросетевая модель, позволяющая проводить классификацию участков звукового сигнала на классы «шум»/«речь», лежащая в основе алгоритма обнаружения речевой активности во входном звуковом сигнале для отделения человеческой речи от фонового шума или тишины.

Обзор алгоритмов VAD

Реализация алгоритмов VAD базируется на следующих положениях относительно речевого сигнала:

- речь является нестационарным сигналом, плавно изменяющимся во времени, т.е. на коротких отрезках времени длиной 20 – 30 мс артикуляционные органы речи не успевают перестроиться, поэтому в пределах фрейма с такой длиной речевой сигнал считается однородным;
- фоновый шум обычно стационарен на более длинном отрезке времени, немного изменяясь со временем;

- в сигналах, содержащих разборчивую речь, уровень речевого сигнала обычно выше уровня фонового шума;
- вокализованные участки речи имеют квазипериодическую структуру из-за наличия в них периода основного тона – низкой частоты, создаваемой речевым трактом человека при смыкании и размыкании связок, в отличие от смычных и шумных глухих звуков;
- невокализованные фонемы (шумные глухие щелевые или смычно-щелевые звуки *с*, *ч*, *ш*, *ф* и др.) представляют собой речевые участки с сильным высокочастотным шумом (звуки).
- частота основного тона имеет большую амплитуду по сравнению с другими частотами в речевом спектре и варьируется от 80 Гц до 180 Гц для мужчин и от 160 Гц до 260 Гц для женщин [1], что соответствует низкочастотной области сигнала, поэтому наличие речи определяется именно по этой части спектра.

Во всех VAD сигнал разбивается на фреймы, как правило, с перекрытием, длиной 20 – 30 мс. Характеристики сигнала, по которым определяется наличие речевой составляющей, вычисляются для каждого фрейма, после чего принимается решение, является ли фрейм речевым сигналом.

Самым простым признаком обнаружения речевой активности является частота пересечения нуля (*Zero Crossing Rate, ZCR*), которая является самой простой мерой частотных свойств сигналов. В плане восприятия значение ZCR соответствует общей оценке тембра звучания по шкале: высокая/низкая частота, глухой/звонкий, шипящий/свистящий.

Кривая ZCR зачастую используется при сегментации речевого сигнала на широкие фонетические классы с помощью порога, определяемого для отсечения вокализованных участков, для которых значение ZCR будет существенно ниже, чем для невокализованных фонем.

Для определения порога по ZCR необходимо в сигнале выделить тональные и высокочастотные шумовые участки и определить максимальное значение частоты пересечения нуля на тональных участках. Это значение используется в качестве порога для отсечения шума, т.е. те участки сигнала, на которых значение частоты пересечения нуля превышает заданный порог, считаются шумовыми, не содержащими речь.

Одним из распространенных и простых признаков наличия речи в сигнале является его энергия – характеристика для обнаружения речевой активности во временной области сигнала. Самые простые алгоритмы VAD основаны на вычислении энергии. В идеальном случае, без наличия шумов, на участках пауз энергия сигнала близка к нулю. Если энергия сигнала поднимается на пороговое значение выше уровня шума, то предполагается, что увеличение энергии связано с голосом.

Краткосрочная мощность (*Short Term Power, STP*) вычисляется по следующей формуле:

$$STP(i) = \frac{1}{N}\sum_{n=0}^{N-1} x_n^2(i),$$

где $x_n(i)$, n = 0, 2, … , N-1 – последовательность из N отсчетов сигнала для -го фрейма.

В данном случае для обнаружения речи выдвигается предположение, что ее компоненты демонстрируют более высокие значения мощности по сравнению с фоновым шумом. Такое допущение об увеличении мощности оправдано из-за эффекта Ломбарда [2], согласно которому говорящий повышает голос в шумной обстановке.

Поэтому, если значение STP выше некоторого порога, то участок сигнала относится к речевой активности. Фиксированное пороговое значение требует априорных знаний об уровнях шума и речи, однако, как правило, уровень шума в большинстве приложений заранее неизвестен, поэтому он должен оцениваться во время разговора из предположения, что первые фреймы сигнала не содержат речь. Нормализация мощности увеличивает разделимость между компонентами речи и шума, однако нестационарные помехи, такие как ударные шумы, вызывают ложные срабатывания детектора речевой активности на основе энергетических характеристик. Кроме того, при наличии высокоамплитудного фонового шума точность систем, использующих энергию для определения границ речи, сильно падает.

В рассмотренных характеристиках наличия речи для выделения интервалов активности и пауз есть еще один существенный недостаток, связанный с потерей невокализованных звуков (принятие их за шум) в силу низкого значения их мощности, которое незначительно отличается от мощности шума (паузы).

Наличие фоновых шумов и большая вероятность принять невокализованные звуки за шум осложняют использование ZCR и STP в качестве параметра для VAD-алгоритма и являются причиной применения более сложных методов, которые связаны с дискретными спектральными преобразованиями, отображающими сигнал из временной области в частотную. Самое полезное свойство представления сигнала в частотной области заключается в том, что отличительные характеристики речевой активности сохраняются и при наложении белого шума.

Учитывая вышеперечисленные особенности речи, эксперты выделяют несколько частотных признаков.

Спектральная мера плоскостности (*Spectral Flatness Measure, SFM*) является показателем «шумности» спектра и её можно использовать для того, чтобы различить речевую и шумовую активности:

$$SFM(i) = \frac{\left(\prod_{n=1}^{N-1} X_n(i)\right)^{1/N}}{\frac{1}{N}\sum_{n=0}^{N-1} X_n(i)},$$

где $X_n(i)$, $n = 1, 2, \ldots, N$ – спектр сигнала из N частот после преобразования Фурье для -го окна.

При значениях SFM, близких к нулю, можно утверждать, что сигнал имеет несколько мощных гармоник, и если они сосредоточены в области низких частот, то можно предположить, что это голос. В обратном случае, при близости данного коэффициента к единице, можно предположить, что данный сигнал близок к белому шуму, для которого спектр частот распределен равномерно [3].

Другим признаком речевой активности является коэффициент полосы частот спектра (*Spectrum Frequency Band Ratio*), который представляет собой отношение суммы амплитуд определенной полосы частот к сумме всех амплитуд спектра [4], [5]. Он вычисляется по формуле:

$$SFBR(i) = \frac{\sum_{n=T}^{K} X_n(i)}{\sum_{n=0}^{N-1} X_n(i)},$$

где T и K – границы диапазона частот, которые в случае присутствия в сигнале речевой активности должны иметь наибольшую мощность. При исследованиях используют значения T и K равными 80 и 1000 Гц соответственно, где сосредоточена основная энергия спектра вокализованных звуков.

Разбиение сигнала на частотные диапазоны используется алгоритмом WebRTC VAD [6], хорошо зарекомендовавшим себя решением от Google. WebRTC VAD разработан для предоставления браузерам и мобильным приложениям функций связи в реальном времени с помощью простых API. Алгоритм анализирует участки аудиозаписи длиной 10, 20 или 30 мс на шести частотных диапазонах: 80 – 250 Гц, 250 – 500 Гц, 500 – 1000 Гц, 1000 – 2000 Гц, 2000 – 3000 Гц, 3000 – 4000 Гц, на каждом из которых вычисляется энергия спектра Фурье, а также общая энергия всего спектра. Если общая энергия выше некоторого порогового значения, то для каждого диапазона вычисляется отношение правдоподобия того, что он содержит речь с помощью модели гауссовых смесей, обученных на два класса – шум и речь, после чего принимается решение по пороговому значению.

Работа VAD в узкополосном речевом кодеке G.729, который используется в телефонной связи в сети Интернет (VoIP) направлена на то, чтобы паузы, которые содержит речевой сигнал, не передавать по каналу, а заменить «комфортным шумом», который генерируется для слушателя. Кодек G.729 принимает сигнал длительностью 10 мс, и вычисляет для него значения четырех параметров: разность энергий всего диапазона частот; разность энергий диапазона низких частот; искажение спектра; разность частоты переходов через ноль. После этого принимается решение о наличии или отсутствии голосовой активности [7].

На сегодняшний день существует множество подходов к реализации VAD, среди них все большую популярность набирают решения на основе методов глубокого обучения.

До появления парадигмы глубокого обучения качество алгоритмов машинного обучения в задачах обработки и распознавания речи сильно зависело от способов формирования признакового пространства. Разработчики вынуждены были тратить много усилий на исследование сложных акустических признаков. Ситуация изменилась с появлением глубоких нейронных сетей. Свойство скрытых слоев глубокой сети извлекать из входного набора данных более абстрактные признаки [8] дало возможность автоматически выводить более высокие абстракции из простых спектральных представлений и использовать глубокое обучение при реализации VAD.

В силу того, что отображение акустического сигнала в частотную область и представляется в виде спектрограммы, следует использовать разновидность глубоких сетей, которая называется свёрточной нейронной сетью (СНС). В задачах обработки изображений она позволяет добиться значительного улучшения точности и снижения вычислительной сложности по сравнению с полносвязными сетями [9].

К алгоритмам, использующим методы глубокого обучения, относится VadNet [10]. Модель принимает в качестве входных данных спектрограммы последовательности фреймов необработанного аудиосигнала, которые передаются через три сверточных слоя. Полученные карты признаков обрабатываются двухслойной рекуррентной сетью. Последний слой является полносвязным, к выходу которого применяется функция Softmax для получения вероятности отнесения входного сигнала к речи или шуму.

К недостаткам данного подхода следует отнести вычислительную сложность из-за значительного размера входа и использования рекуррентной нейронной сети.

В данной работе для бинарной классификации фрейма входного сигнала на «речь»/«шум» предлагается модель сверточной сети, имеющей значительно меньше параметров, входными данными для которой является изображение спектрограммы.

Формирование dataset для обучения и тестирования моделей, используемых VAD, включая нейросетевую, – задача, которую необходимо решить на раннем этапе разработки VAD. Особенно важную роль подбор dataset играет для обучения глубокой нейронной сети.

Описание обучающего набора данных

Для создания обучающей и тестовой выборки создан размеченный аудиокорпус на основе нескольких наборов данных, находящихся в открытом доступе. Эти наборы включают разные виды шумовых эффектов и речевые данные на разные языках (в рамках решаемой задачи язык речевого сообщения не имеет значения).

Использовались следующие наборы:

1) *FSDnoisy18k* [11] – набор аудиоданных, собранный для исследования различных типов шумов при классификации звуковых событий. Он содержит записи общей продолжительностью 42,5 часа, из которых 2,4 часа представляют собой обучающий набор данных, содержащих незашумленную речь, а остальные данные – зашумленные звуковые сигналы, распределенные по 20 классам шумов, среди которых шумы окружающей среды (улицы, шагов и офисного помещения), а также музыкальные шумы.

Для обучения модели в работе из FSDnoisy18k использовалась выборка, суммарно по всем классам шумов состоящая из 17 585 записей общей длительностью около 40 часов, для тестирования – из 947 записей общей длительностью 2,5 часов.

2) *Microsoft Scalable Noisy Speech Dataset (MS-SNSD)* – расширяемый набор аудиоданных от Microsoft [12]. MS-SNSD предназначен для обучения моделей глубокой нейронной сети с целью подавления фонового шума и содержит большую коллекцию файлов с чистой английской речью и файлов с разнообразным шумом окружающей среды в формате .wav, с частотой дискретизации 16 000 Гц.

Для обучения сети из MS-SNSD использовались только файлы, содержащие чистую речь, обучающая выборка составила 23 075 записей общей длительностью 32 часов, тестовая – 1 100 записей общей длительностью 1,6 часов.

Для обработки аудиоданных из описанных корпусов с целью преобразования тестовой и обучающей выборок в спектрограммы использовались следующие параметры:

- частота дискретизации 16 000 Гц;
- глубина квантования – 32 бит;
- размер фрейма – 512 отсчетов;
- размер перекрытия – 256 отсчетов.

Полученные спектрограммы представляют собой изображения в формате RGB размером 64×64 пикселей.

Описание модели сверточной сети для определения речевой активности

Модель проектировалась на основе экспериментальных исследований, посредством подбора количества слоёв и параметров на каждом из них. Модель с лучшими показателями на экспериментальной выборке имеет следующие параметры:

- размер входных векторов – 64×64;
- оптимизатор градиентного спуска – Adam;
- функция потерь – кросс-энтропия;
- количество эпох – 10;

– функция активации слоев – *rectified linear unit (ReLU)*;
– количество сверточных слоев – 6 свёрточных слоёв (*Conv*);
– размер ядра на всех слоях свёртки – 3×3;
– размер батча – 32;
– операция подвыборки – MaxPooling;
– коэффициенты dropout-регуляризации 0,25 и 0,5;

Архитектура сверточной сети изображена на рис. 1.

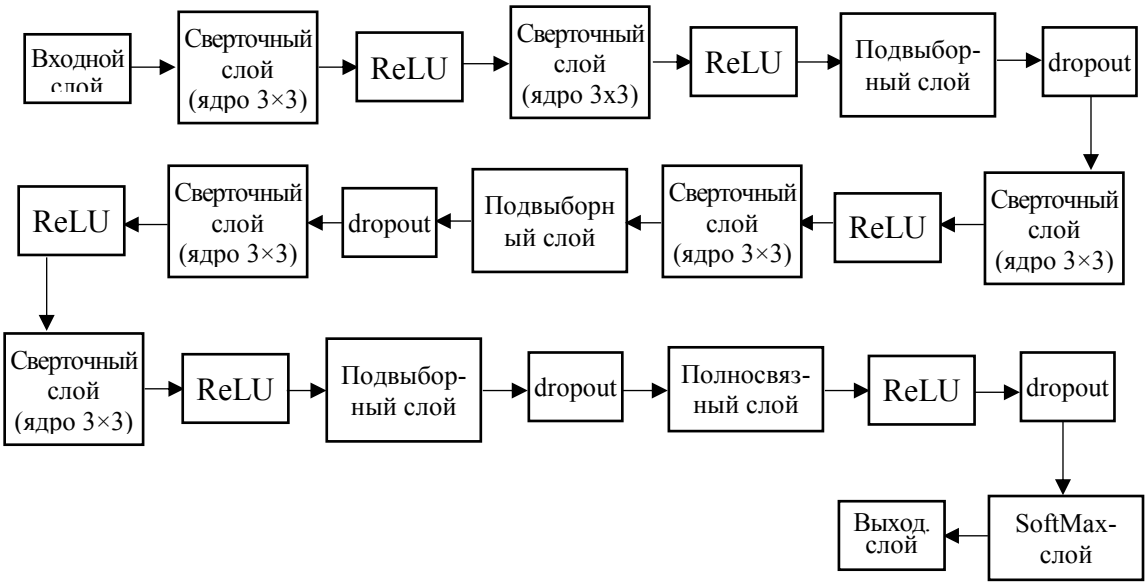

Рисунок 1 – Архитектура сверточной сети для классификации фреймов сигнала

Для оценки качества модели использовались точность, полнота и F1-мера [13]. В табл. 1 приведены результаты проверки качества модели на тестовой выборке.

Таблица 1 – Результаты тестирования качества предложенной модели

Класс	Метрика оценки качества модели		
	Точность	Полнота	F1-мера
Шум	0,91	0,87	0,89
Речь	0,98	0,80	0,88

Исходя из результатов тестирования, можно сделать вывод о высокой эффективности предложенной архитектуры, обученной на различных видах фоновых шумов и имеющей точность более 90%.

Выводы

Основным недостатком существующих VAD-алгоритмов является недостаточная точность и значительное ухудшение точности определения границ речевых сегментов при наличии фоновых шумов, вызванных разнообразием внешних факторов, влияющих на запись и передачу голоса. Кроме того, существует большая вероятность принять невокализованные звуки за шум, а высокоамплитудный шум – за речь.

Благодаря свойству скрытых слоев глубокой нейронной сети извлекать из входного набора данных более абстрактные признаки, на сегодняшний день СНС, на вход которым подаются изображения спектрограмм, являются наиболее эффективным классификатором фреймов аудиосигнала на «шум»/«речь». В данной работе для

бинарной классификации фрейма входного сигнала предлагается модель сверточной сети, использующей относительно немного параметров, но обладающей достаточной точностью (более 90%) после ее обучения на базе корпусов FSDnoisy18k и MS-SNSD, которые содержат различиные типы шумов.

Список литературы

1. Голосовая биометрия в сфере VoIP [Электронный ресурс]. – URL: https://www.itworld.ru/tech/science/142282.html (дата обращения: 16.07.2019).
2. J-C. Junqua. The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex [Текст] / J-C. Junqua // Speech Communication. – Vol. 20(1), 1996. – P. 13–22.
3. Moattar Mohammad. A simple but efficient real-time voice activity detection algorithm [Электронный ресурс] / Moattar Mohammad, Homayoonpoor Mahdi // 17th European Signal Processing Conference (EUSIPCO 2009) Glasgow, Scotland, August 24-28, 2009. – URL: https://www.researchgate.net/publication/255667085_A_simple_but_efficient_real-time_voice_activity_detection_algorithm (дата обращения: 16.02.2022).
4. Voice Activity Detection for Voice User Interface [Электронный ресурс]. – URL: https://medium.com/linagoralabs/voice-activity-detection-for-voice-userinterface-2d4bb5600ee3 (дата обращения: 16.02.2022).
5. Simon Graf, Tobias Herbig, Markus Buck, Gerhard Schmidt. Features for voice activity detection: a comparative analysis [Электронный ресурс] // EURASIP Journal on Advances in Signal Processing, 2015. – URL: https://asp-eurasipjournals.springeropen.com/track/pdf/10.1186/s13634-015-0277-z.pdf (дата обращения: 16.02.2022).
6. Python interface to the WebRTC Voice Activity Detector [Электронный ресурс]. – URL: https://github.com/wiseman/py-webrtcvad (дата обращения: 16.02.2022).
7. Panji Setiawan, Stefan Schandl et al. On the ITU-T G.729.1 Silence compression scheme [Электронный ресурс] // 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008. – URL: https://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569104920.pdf (дата обращения: 16.02.2022).
8. Гудфеллоу Я. Глубокое обучение [Текст] / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; пер. с анг. А. А. Слинкина. – 2-е изд. – М. : ДМК Пресс, 2018. – 652 с.
9. Николенко С. Глубокое обучение [Текст] / С. Николенко, А. Кадурин, Е. Архангельская. – СПб. : Питер, 2018. – 480 с.
10. Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant? [Электронный ресурс] / Wagner Johannes, Schiller Dominik, Seiderer Andreas, Andre, Elisabeth 10.21437/ Interspeech, 2-6 September 2018, Hyderabad, 2018. – P. 147–151. – URL: https://d-nb.info/1202249140/34 (дата обращения: 16.02.2022).
11. Learning Sound Event Classifiers from Web Audio with Noisy Labels [Электронный ресурс] / E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, X. Serra // arXiv.org. 2019. – URL: https://arxiv.org/pdf/1901.01189.pdf (дата обращения: 16.02.2022).
12. A scalable noisy speech dataset and online subjective test framework [Электронный ресурс] / Chandan K. A. Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, Johannes Gehrke. // arXiv.org. 2019. – URL: https://arxiv.org/ftp/arxiv/papers/1909/1909.08050.pdf (дата обращения: 16.02.2022).
13. Луценко Е. В. Нечеткое мультиклассовое обобщение классической F-меры достоверности моделей Ван Ризбергена в АСК-анализе и системе «Эйдос» [Электронный ресурс] / Е. В. Луценко // Научный журнал КубГАУ. – 2016. – № 123. – URL: https://cyberleninka.ru/article/n/nechetkoe-multiklassovoe-obobschenie-klassicheskoy-f-mery-dostovernosti-modeley-van-rizbergena-v-ask-analize-i-sisteme-eydos (дата обращения: 16.02.2022).
14. Харламов А. А. Анализ текстов: лингвистика, семантика, прагматика в рамках когнитивного подхода [Текст] / А. А. Харламов, Т. В. Ермоленко // Проблемы искусственного интеллекта. – 2015. – № 0 (1). – С. 106–115.

References

1. *Golosovaya biometriya v sfere* VoIP [Golosovaya biometriya v sfere VoIP] [Elektronnyj resurs], URL: https://www.itworld.ru/tech/science/142282.html (data obrashcheniya: 16.07.2019).
2. J-C Junqua. The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*. Vol. 20(1), 1996, P. 13–22.
3. Moattar Mohammad, Homayoonpoor Mahdi. A simple but efficient real-time voice activity detection algorithm [Elektronnyj resurs]. *17th European Signal Processing Conference (EUSIPCO 2009)* Glasgow, Scotland, August 24-28, 2009,

URL: https://www.researchgate.net/publication/255667085_A_simple_but_efficient_real-time_voice_activity_detection_algorithm (data obrashcheniya: 16.02.2022).
4. *Voice Activity Detection for Voice User Interface* [Elektronnyj resurs], URL: https://medium.com/linagoralabs/voice-activity-detection-for-voice-userinterface-2d4bb5600ee3 (data obrashcheniya: 16.02.2022).
5. Simon Graf, Tobias Herbig, Markus Buck, Gerhard Schmidt. Features for voice activity detection: a comparative analysis [Elektronnyj resurs]. *EURASIP Journal on Advances in Signal Processing*, 2015, URL: https://asp-eurasipjournals.springeropen.com/track/pdf/10.1186/s13634-015-0277-z.pdf (data obrashcheniya: 16.02.2022).
6. *Python interface to the WebRTC Voice Activity Detector* [Elektronnyj resurs]. URL: https://github.com/wiseman/py-webrtcvad (data obrashcheniya: 16.02.2022).
7. Panji Setiawan, Stefan Schandl et al. On the ITU-T G.729.1 Silence compression scheme [Elektronnyj resurs]. *16th European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, August 25-29, 2008. URL: https://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569104920.pdf (data obrashcheniya: 16.02.2022).
8. Gudfellou YA., Bendzhio I., Kurvill' A. *Glubokoe obuchenie* [Deep learning] ; per. s ang. A. A. Slinkina, 2-e izd., M., DMK Press, 2018, 652 s.
9. Nikolenko S., Kadurin A., Arhangel'skaya E. *Glubokoe obuchenie* [Deep learning] ; SPb., Piter, 2018; 480 s.
10. Wagner Johannes, Schiller Dominik, Seiderer Andreas, Andre, Elisabeth. Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant? [Elektronnyj resurs] 10.21437/ Interspeech, 2-6 September 2018, Hyderabad, 2018. – P. 147-151. – URL: https://d-nb.info/1202249140/34 (data obrashcheniya: 16.02.2022).
11. E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, X. Serra. Learning Sound Event Classifiers from Web Audio with Noisy Labels [Elektronnyj resurs]. *arXiv.org*. 2019, URL: https://arxiv.org/pdf/1901.01189.pdf (data obrashcheniya: 16.02.2022).
12. Chandan K. A. Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework [Elektronnyj resurs] // arXiv.org. 2019. – URL: https://arxiv.org/ftp/arxiv/papers/1909/1909.08050.pdf (data obrashcheniya: 16.02.2022).
13. Lucenko E. V. Nechetkoe mul'tiklassovoe obobshchenie klassicheskoj F-mery dostovernosti modelej Van Rizbergena v ASK-analize i sisteme «Ejdos» [Elektronnyj resurs] // Nauchnyj zhurnal KubGAU. 2016. №123. – URL: https://cyberleninka.ru/article/n/nechetkoe-multiklassovoe-obobschenie-klassicheskoy-f-mery-dostovernosti-modeley-van-rizbergena-v-ask-analize-i-sisteme-eydos (data obrashcheniya: 16.02.2022).
14. Kharlamov A. A., Ermolenko T. V. Analiz tekstov: lingvistika, semantika, pragmatika v ramkakh kognitivnogo podkhoda [Text analysis: linguistics, semantics, pragmatics within the cognitive approach]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2015, no. 0(1), pp. 106–115.

RESUME

I. N. Savenkov, T. V. Yermolenko, A. V. Tsybik

Developing a VAD-algorithm based on deep learning

The article provides a brief overview of the methods of constructing VAD algorithms, as well as the main characteristics that determine the presence of speech in the audio signal. The main disadvantage of the existing speech detection algorithms based on the energy characteristics of the signal is that frames containing noisy deaf slit or bow-slit sounds can be classified as noise, and frames containing high-amplitude noise can be classified as speech.

The use of deep neural networks to classify signal frames into "noise"/"speech" classes made it possible to automatically derive higher abstractions from simple spectral representations in hidden layers and use deep learning in the implementation of VAD.

For binary classification, a convolutional network model is proposed, the input data for which is an image of a Fourier spectrogram in RGB with a size of 64×64 pixels. The proposed architecture contains six convolutional layers and two fully connected layers.

The training and testing of the model was carried out on a marked audio corpus formed on the basis of the FSDnoisy18k and MS-SNSD datasets, which contain both recordings of pure speech and recordings with various musical and environmental noises distributed across 20 classes.

Testing has shown the high efficiency of the model, which allows classifying frames of a signal containing speech with an accuracy of more than 90%.

РЕЗЮМЕ

И. Н. Савенков, Т. В. Ермоленко, А. В. Цыбик
Разработка VAD-алгоритма на основе глубокого обучения

В статье дан краткий обзор методов построения VAD-алгоритмов, а также основных характеристик, определяющих наличие речи в аудиосигнале. Основным недостатком существующих алгоритмов детектирования речи, основанных на энергетических характеристиках сигнала, является то, что фреймы, содержащие шумные глухие щелевые или смычно-щелевые звуки, могут классифицироваться как шум, а фреймы, содержащие высокоамплитудный шум, – как речь.

Использование глубоких нейросетей для классификации фреймов сигнала на классы «шум»/«речь» дало возможность автоматически выводить более высокие абстракции из простых спектральных представлений в скрытых слоях и использовать глубокое обучение при реализации VAD.

Для бинарной классификации в работе предложена модель сверточной сети, входными данными для которой является изображение спектрограммы Фурье в RGB размером 64×64 пикселей. Предложенная архитектура содержит шесть сверточных слоев и два полносвязных слоя.

Обучение и тестирование модели проводилось на размеченном аудиокорпусе, сформированном на основе наборов данных FSDnoisy18k и MS-SNSD, которые содержат как записи чистой речи, так и записи с разнообразными музыкальными шумами и шумами окружающей среды, распределенные по 20 классам.

Тестирование показало высокую эффективность модели, позволяющей классифицировать фреймы сигнала, содержащего речь, с точностью более 90%.

Статья поступила в редакцию 27.12.2021.