УДК 004.93

КОМБИНИРОВАННЫЙ ДЕТЕКТОР ГОЛОСОВОЙ АКТИВНОСТИ

Стефаниди А.Ф., *аспирант Ярославского государственного университета им. П.Г. Демидова, e-mail: antonstefanidi@mail.ru*
Приоров А.Л., *д.т.н., профессор Ярославского государственного университета им. П.Г. Демидова, e-mail: andcat@yandex.ru*
Топников А.И., *к.т.н., Ярославский государственный университет им. П.Г. Демидова, e-mail: topartgroup@gmail.com*
Хрящев В.В., *к.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: vhr@yandex.ru*

CASCADE VOICE ACTIVITY DETECTOR

Stefanidi A.F., Priorov A.L., Topnikov A.I., Khryashchev V.V.

The problem of analyzing speech signals is considered. This type of signals consists of speech, external noise, recording device noise and pauses. The presence of pauses, noise and interference from the point of view of voice biometrics systems is a negative factor affecting the accuracy of personality recognition.The aim of the work is to develop a voice activity detector to improve the accuracy of speech fragments selection.

The original VADSpeakersDB dataset has been prepared, containing 138,000 fragments of Russian-language speech, noises and pauses. A combined voice activity detector (CDGA) has been developed and tested. The solution has a high accuracy of detecting speech fragments – above 90 %. The accuracy of determining fragments of voice activity when using CDGA increases by 2-3 % in comparison with the analogues considered in the work.

The detector can be used to process speech signals in the task of biometric identification. The VADSpeakersDB dataset can be used to develop and test solutions in the field of speech signal processing that are of practical interest to the domestic market.

Key words: digital speech processing, voice activity detection, stacking, machine learning, random forest.

Ключевые слова: цифровая обработка речевых сигналов, детектирование голосовой активности, стекинг алгоритмов, машинное обучение, ансамбль решающих деревьев.

Рассматривается задача анализа речевых сигналов. Данный тип сигналов состоит из речи, внешних шумов, шумов записывающего устройства и пауз. Наличие пауз, шумов и помех с точки зрения систем голосовой биометрии является негативным фактором, влияющим на точность распознавания личности. Целью работы является разработка детектора голосовой активности для повышения точности выделения речевых фрагментов.

Подготовлен оригинальный набор данных VADSpeakersDB, содержащий 138 000 фрагментов русскоязычной речи, шумов и пауз. Разработан и протестирован комбинированный детектор голосовой активности (КДГА). Решение имеет высокую точность детектирования речевых фрагментов – выше 90 %. Точность определения фрагментов голосовой активности при использовании КДГА повышается на 2-3 % в сравнении с рассмотренными в работе аналогами. Детектор может применяться для обработки речевых сигналов в задаче биометрической идентификации. Набор данных VADSpeakersDB может быть использован для разработки и тестирования решений в области обработки речевых сигналов, имеющих практический интерес для отечественного рынка.

Введение

В настоящее время большой интерес в области анализа речевых сигналов получила задача распознавания диктора [1-3]. Алгоритмы голосовой биометрии применяются, например, при построении систем контроля управления доступом, в банкинг приложениях для верификации клиентов. Методы распознавания личности по голосу используются в криминалистике для борьбы с телефонным терроризмом. Однако такого рода подходы имеют сильную зависимость от качества анализируемых речевых сигналов. Наличие пауз, шумов и помех ухудшает точность работы методов идентификации диктора [4-7], поэтому важно начинать анализ фонограммы с выделения фрагментов, содержащих речь. Для этого используются детекторы голосовой активности (ДГА, Voice Activity Detector, VAD), способные разделять исходный сигнал на речевые и неречевые участки [5]. Эти алгоритмы, с одной стороны, должны быть достаточно точными, а с другой, – достаточно простыми, так как обычно они являются лишь одним из этапов предобработки в составе систем распознавания.

Цель работы – разработать оригинальный составной детектор на основе стекинг-объединения набора из независимых детекторов голосовой активности.

Классические методы детектирования голосовой активности

В настоящее время известно большое количество детекторов речевой активности, от самых простых решений до алгоритмов на основе нейронных сетей и других методов машинного обучения. Однако применение современных подходов позволяет не только повысить точность, но и приводит к повышению вычислительной

сложности решения, что не всегда уместно [8-9]. Поэтому сконцентрируемся в этом исследовании на алгоритмах, требующих относительно небольшого объема вычислений.

Энергия является одной из наиболее простых и важных характеристик аудиосигналов. В теории сигналов энергия является количественной характеристикой, отражающей определенные свойства сигнала и динамику изменения его значений во времени, в пространстве или по некоторым другим аргументам. Она определяется как квадрат функции амплитуды сигнала. Также энергия может быть рассчитана как интеграл от мощности по всему интервалу существования рассматриваемого сигнала.

Одним из самых простых видов ДГА является алгоритм на основе анализа энергии (далее – ДГА1) [10-11]. Сигнал во временной области делится на окна длиной 10-30 мс. Далее, для каждого окна вычисляется сумма квадратов амплитуды. После этого проводится пороговая обработка. Если значение энергии больше заданного порога θ, то фрагмент оставляют. Логика этого процесса определяется тем, что речевые фрагменты имеют высокий уровень энергии, тогда как фрагменты, содержащие шум/паузы, обладают, как правило, меньшей энергетикой. Исключением из такой логики является импульсный шум. Алгоритм на основе анализа энергии математически можно описать следующим образом:

$$S = \{\vec{S}_1, \vec{S}_2, \vec{S}_3 ..... \vec{S}_j\}, \text{ где } \vec{S}_j = (s_1, s_2, s_3 ..... s_N),$$

$$E_j = \sum_{i=1}^{N} E(i) = \sum_{i=1}^{N} s^2(i),$$

$$\vec{V} = \begin{cases} \vec{S}_j, \text{ если } \theta <= E_j \\ 0, \text{ если } \theta > E_j, \end{cases}$$

$$S' = \{\vec{V}_1, \vec{V}_2, \vec{V}_3, ..., \vec{V}_w\},$$

где S – исходный речевой сигнал, $\vec{S}_j$ – j-й фрагмент исходного сигнала, $s(i)$ – амплитуда i-го отсчета, $E(i)$ – энергия i-го отсчета, E_j – энергия j-го фрагмента исходного сигнала, N – длина окна, $\vec{V}$ – речевой фрагмент, w – количество окон, содержащих речь, S' – обработанный сигнал [10].

Значение порога θ определяется следующим образом. Для каждого фрагмента фонограммы $\vec{S}_j$ высчитывается энергия, после чего определяется минимальное и максимальное значение энергии для всей фонограммы. Затем эмпирически подбирается коэффициент k. Регулируя его, можно контролировать порог θ, рассчитываемый следующим образом:

$$\theta = k \cdot (E_{\max} - E_{\min}),$$

где $E_{\max}$, $E_{\min}$ – максимальное и минимальное значение энергии фонограммы.

Другим возможным методом построения ДГА является подход на основе анализа энергии Тигера-Кайзера (далее – ДГА2) [10-11]. Принцип работы алгоритма основывается на том, что сигнал не разбивается на окна, а для каждого временного отсчета вычисляется энергия следующим образом:

$$E(i) = s^2(i) - s(i-1)s(i+1),$$

где $s(i)$ – i-й отсчет фонограммы.

Также методы анализа фонограмм могут реализовываться с использованием частотного представления сигнала на основе дискретного преобразования Фурье (ДПФ). Для проведения исследований в работе реализован алгоритм ДГА на основе частотного анализа фонограмм (далее – ДГА3). Опишем подробно основные этапы работы такого детектора. Вначале исходная фонограмма обрабатывается с помощью ДПФ, что в результате формирует спектрограмму. Следующим этапом вычисляется периодограмма, которая определяется как модуль квадрата спектрограммы. Затем периодограмма обрабатывается полосовым фильтром, поскольку речь человека составляет ограниченный диапазон частот от 300 Гц до 3,4 кГц. На завершающем этапе обработки для каждого фрагмента определяется максимальное значение величины мощности спектра. Если значение устанавливается выше определяемого порога θ, то фрагмент помечается как участок фонограммы, содержащий исключительно голос диктора [12].

Рис. 1. Структурная схема предложенного алгоритма КДГА

Комбинированный детектор голосовой активности

Так как рассмотренные выше алгоритмы не требовательны к вычислительным ресурсам, можно попробовать объединить их в единый алгоритм с помощью стекинга для повышения итоговой точности детектирования [13]. Основная идея стекинга состоит в том, чтобы использовать каждый ДГА в качестве независимого детектора, а их выходы объединить для последующего анализа обобщающим классификатором. На рис. 1 представлена схема предложенного решения – комбинированный детектор голосовой активности (далее – КДГА).

На вход алгоритма поступает фрагмент фонограммы длительностью в 10 мс. Далее каждый из детекторов в каскаде анализирует входной образец. После этого формируются 3 независимые предсказания, которые подаются на вход обобщающего классификатора. Последний принимает итоговое решение, к какому из классов отнести входной фрагмент фонограммы – «речь» или «не речь». В качестве метамодели используется классификатор на основе ансамбля решающих деревьев [14-15].

Подготовленный набор речевых сигналов

Для проведения исследования собран собственный уникальный набор речевых сигналов VADSpeakersDB. Набор представляет собой запись живой русскоязычной речи 23 дикторов с частотой дискретизации 16 кГц. Речевые данные являются достаточно сбалансированными по гендерному признаку – 55 % составляет речь мужчин и 45 % – речь женщин. Каждого диктора записывали в течение 60 секунд. В итоге общая длительность набора VADSpeakersDB составила 23 минуты. В качестве программного обеспечения использовалось приложение Zoom, поскольку в настоящее время это одно из самых распространенных решений для удаленного взаимодействия, что в некоторой степени упрощает организацию процесса сбора данных. Важно отметить, что существует потребность в создании алгоритмов биометрической идентификации, способных качественного работать в условиях видеоконференцсвязи на базе Skype, Zoom, а также ряда других аналогичных приложений.

Далее, данные в ручном режиме размечались специалистами. Для этого весь набор делился на неперекрывающиеся фрагменты длительностью в 10 мс. Каждый фрагмент прослушивался специалистом, который в итоге выставлял метку «речь» или «не речь». В табл. 1 представлены основные характеристики подготовленного набора фонограмм.

Таблица 1. Основные характеристики собранного набора фонограмм VADSpeakersDB

Язык	Русский
Частота дискретизации	16 кГц
Количество дикторов	23
Общая длительность	23 мин.
Длительность одного фрагмента	10 мс
Общее количество фрагментов	138 000
Количество фрагментов класса «речь»	68,28 %
Количество фрагментов класса «не речь»	31,72 %

Важно отметить, что подготовленный набор VADSpeakersDB является аппаратно-независимым, поскольку запись осуществлялась с использованием большого разнообразия технических устройств. Вследствие этого можно сделать вывод, что подготовленные данные имеют высокую степень сходства с реальными условиями эксплуатации. Это свойство подтверждает актуальность и практическую значимость исследования. Дополнительно стоит отметить, что данный набор может быть использован для разработки и тестирования других решений в области обработки речевых сигналов, имеющих практический интерес для отечественного рынка.

Метрики оценки качества работы детекторов

Поскольку алгоритм ДГА решает задачу бинарной классификации, то для оценки качества работы может быть использована такая метрика, как доля правильных ответов (*acc*). В процессе оценки *acc* важно обратить внимание на несбалансированность данных. Для учета данного свойства вводится аналогичная оценка доли правильных ответов для несбалансированных данных (*accb*). Для подсчета данных метрик определим индикатор корректности распознавания фрагмента речевого сигнала:

$$c(x_i) = \begin{cases} 1, y_i = y_i', \\ 0, y_i \neq y_i', \end{cases}$$

где x_i – i-й фрагмент фонограммы, длительностью 10 мс; $c(x_i)$ – индикатор корректности распознавания i-го фрагмента; y_i – целевая метка фрагмента; y_i' – метка фрагмента, определяющая результат работы ДГА. Тогда метрику acc можно определить, как:

$$acc = \frac{\sum_{i=1}^{n} c(x_i)}{n},$$

где n – количество всех фрагментов длительностью 10 мс в рассматриваемом наборе VADSpeakersDB.

Для подсчета доли правильных ответов на несбалансированных данных в задаче бинарной классификации необходимо воспользоваться выражением:

$$accb = \frac{acc_{y_i=1} + acc_{y_i=0}}{2},$$

где $acc_{y_i=1}$ – доля верно детектированных фрагментов, представляющих класс «речь»; $acc_{y_i=0}$ – доля верно детектированных фрагментов, определяющих класс «не речь».

Также для оценки качества работы ДГА воспользуемся гармоническим средним между точностью и полнотой (F-мера, F):

$$F = 2 \cdot \frac{P \cdot R}{P + R},$$

где P – точность (precision), метрика, определяющая ошибки I рода, R – полнота (recall), метрика, определяющая ошибки II рода [16-17].

Ранее отмечалось, что набор VADSpeakersDB обладает дисбалансом в данных. В соответствии с данным свойством определим F-меру на основе макроусредняющего подхода, то есть расчет метрики внутри

каждого класса («речь», «не речь») с последующей нормировкой на общее количество классов:

$$F_{макро} = \frac{F_{y_i=1} + F_{y_i=0}}{2}.$$

Рассмотренные метрики будут использоваться для исследования работы обученного детектора голосовой активности.

Обучение предлагаемого подхода

Для обучения предложенного классификатора необходимо подготовленный набор данных VADSpeakersDB разложить на обучающую и тестовую выборки. Для проведения эксперимента в процессе обучения использовался подход на основе k-блочной перекрестной проверки (k-fold cross validation) [18-19]. Параметр k определяет, на сколько частей будет разбит обучающий набор. Затем на k-1 частей обучается модель, а оставшаяся часть используется в качестве проверочного множества. Обучение повторяется k раз. В итоге каждая из k-частей участвует в проверке, после чего оценочная характеристика усредняется. В исследовании параметр k задавался равным 10.

В табл. 2 показан принцип разделения набора VADSpeakersDB. На тестирование выделялось 15 % примеров от общей суммы фрагментов речевых данных. В процессе обучения и применения перекрестной проверки 10 % данных использовались для оценки работы классификатора.

Таблица 2. Разделение набора данных VADSpeakersDB

-	Обучающая выборка	Тестовая выборка	Общее количество
Фрагментов, шт.	117300	20700	138000

На рис. 2 изображена схема процесса обучения и тестирования разработанного алгоритма КДГА. Классификатор на базе ансамбля решающих деревьев обладает широким набором настраиваемых параметров. Для настройки используются такие показатели, как общее количество деревьев, выбор критерия разделения, максимальная глубина деревьев, количество анализируемых признаков в каждом из узлов для принятия решения. Подбор оптимальных настроек для классификатора выполняется на основе построения сетки параметров [20]. Для этого выполняется перебор возможных значений для каждого из параметров. Подход является достаточно требовательным к временным и вычислительным ресурсам, поскольку количество обучаемых классификаторов может достигать десятков тысяч. После этого определяется оптимальный классификатор в соответствии с целевой метрикой. В работе в качестве такой метрики использовалась F-мера на основе макроусредняющего подхода.

В процессе обучения и подбора параметров классификатора проведен анализ более 2100 моделей, а также выполнен выбор модели, показавшей наилучшие результаты работы на обучающей и проверочной выбор-

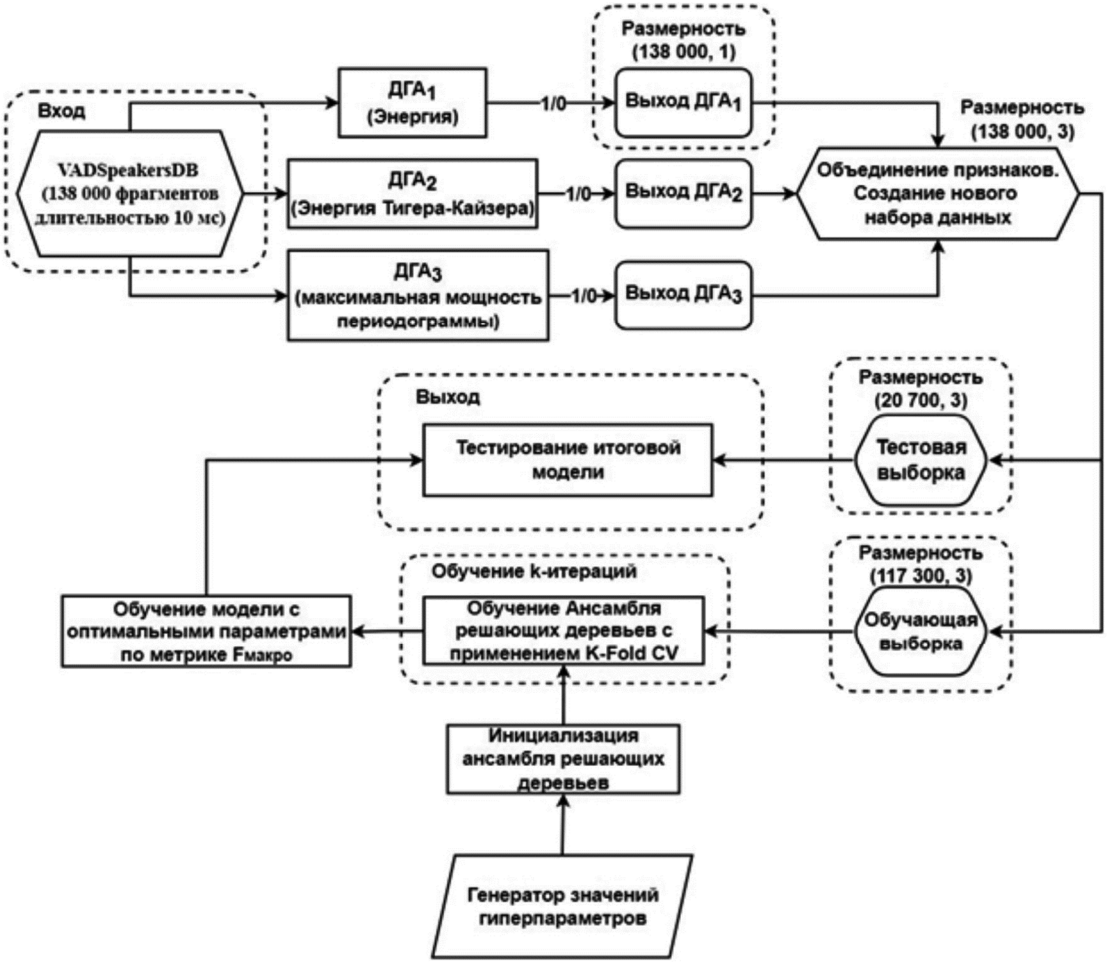

Рис. 2. Схема процесса обучения и тестирования КДГА

ках. На финальном этапе точность классификации проверялась с использованием тестовой подвыборки VADSpeakersDB.

Результаты исследования работы детектора голосовой активности

Для проведения сравнения предварительно подобраны параметры традиционных алгоритмов, отвечающие за выбор порога, так, чтобы максимизировать значения выбранной целевой функции. В табл. 3 представлен сравнительный анализ разработанного алгоритма с классическими подходами. Из полученных результатов видно, что применение КДГА с последующим обучением обобщающего классификатора позволяет повысить точность детектирования голосовых фрагментов на 2-3 %. Улучшение в точности работы обусловлено объединением трёх более «слабых учеников».

Таблица 3. Сравнительный анализ детекторов голосовой активности

Метрики	$ДГА_1$ $(k = 2*10^{-4})$	$ДГА_2$ $(\theta = 3*10^{-6})$	$ДГА_3$ $(\theta = 2*10^{-3})$	КДГА
acc	0,90	0,89	0,89	**0,91**
accb	0,88	0,88	0,87	**0,90**
F	0,93	0,92	0,92	**0,94**
$F_{макро}$	0,88	0,88	0,87	**0,90**

Так прогнозы, полученные на выходе каждого из детекторов, объединяются для анализа «сильным учеником». В процессе обучения модели сформированы весовые параметры для каждого детектора. В итоге каждый ДГА имеет индивидуальный вес при принятии итогового решения. Комбинирование детекторов позволяет повысить точность определения речевых фрагментов.

На рис. 3 изображены результаты работы рассмотренных детекторов. Визуально можно оценить улучшение в выделении зон голосовой активности при использовании алгоритма КДГА.

Заключение

Рассматривался вопрос улучшения качества речевых данных с применением алгоритмов определения фрагментов голосовой активности. Для проведения исследования подготовлен оригинальный набор VADSpeakersDB, который содержит 138 000 фрагментов русскоязычной речи, шумов и пауз. Разработан и протестирован комбинированный детектор голосовой активности КДГА. Принцип его работы основан на объединении независимых простых детекторов с последующим обучением обобщающего классификатора. В качестве базовых ДГА использовались методы на основе анализа энергии, энергии Тигера-Кайзера и спектральной плотности мощности сигнала. Установлено, что решение на базе КДГА имеет высокую точность детектирования речевых фрагментов – выше 90 %. Точность определения фрагментов голосовой активности при его использовании повышается на 2-3 % в сравнении с рассмотренными в работе аналогами. Детектор может применяться для обработки речевых сигналов в задаче биометрической идентификации [21-22].

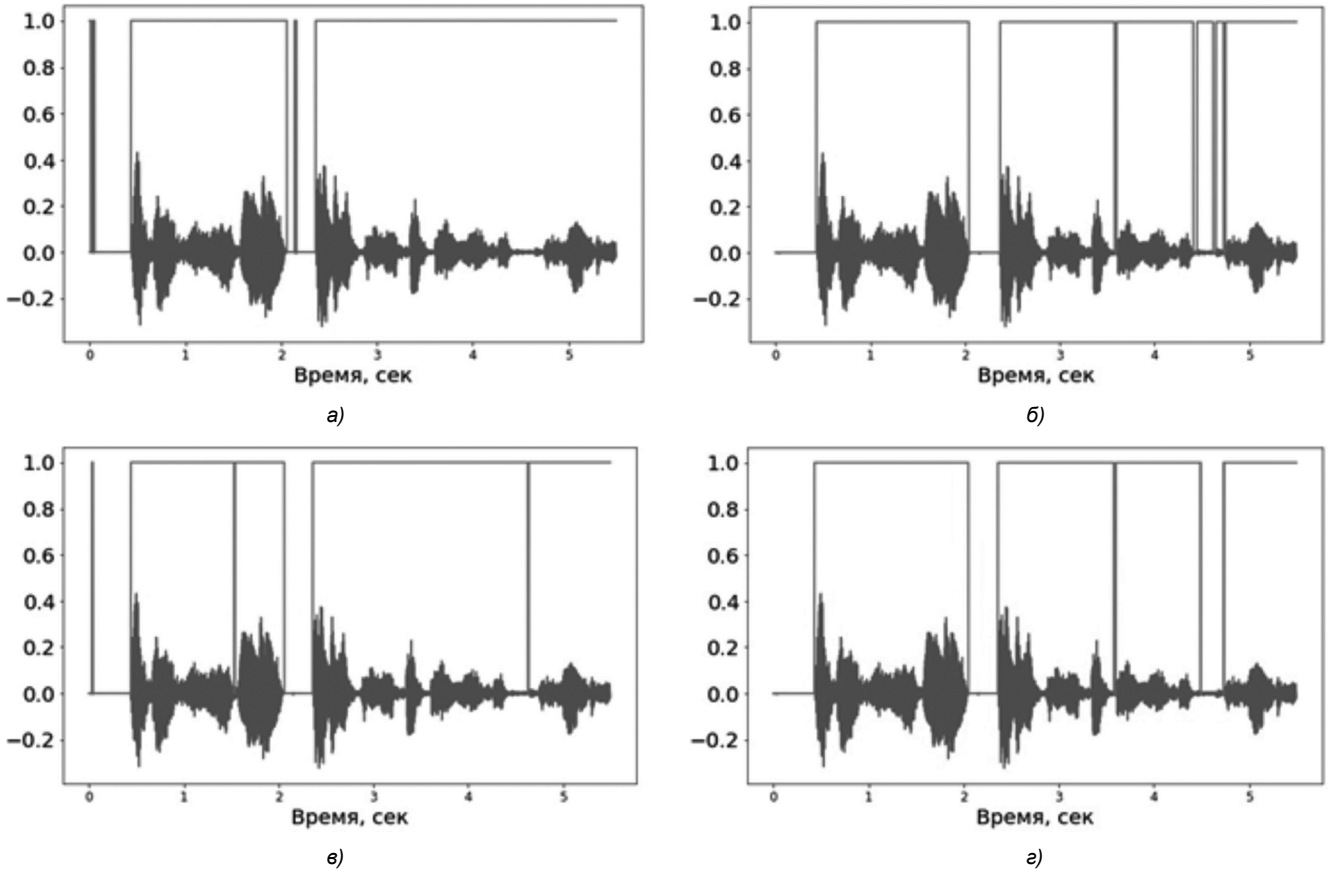

Рис. 3. Результаты работы детекторов голосовой активности: а) на основе анализа энергии; б) на основе анализа энергии Тигера-Кайзера; в) на основе частотного анализа сигнала; г) предложенный алгоритм КДГА

Литература

1. Матвеев Ю.Н. Технология биометрической идентификации личности по голосу и другим модальностям. Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение». 2012. № 3. С. 46-61.

2. Козлов А.В., Кудашев О.Ю., Матвеев Ю.Н., Пеховский Т.С. Система идентификации дикторов по голосу для конкурса NIST SRE 2013. Труды СПИИРАН, 2013. № 2. С. 350-370.

3. Стефаниди А.Ф., Приоров А.Л., Топников А.И., Хрящев В.В. Модификация VGG-архитектуры в задачах унимодальной и мультимодальной биометрии. Цифровая обработка сигналов. 2020. № 3. С. 35-40.

4. Стефаниди А.Ф., Приоров А.Л., Топников А.И., Хрящев В.В. Применение сверточных нейронных сетей в задаче мультимодальной идентификации. Цифровая обработка сигналов. 2020. № 2. С. 52-58.

5. Стефаниди А.Ф., Топников А.И., Приоров А.Л. Использование сверточных нейронных сетей в задаче распознавания диктора. Цифровая обработка сигналов и ее применение (DSPA-2020): докл. 22-й междунар. конф. Москва, 2020. С. 642-646.

6. Lavrentyeva G., Novoselov S., Volokhov, V et al. STC Speaker Recognition System for the NIST SRE 2021. Processing The Speaker and Language Recognition Workshop (Odyssey 2022). 2022, pp. 354-361.

7. Gusev A., Volokhov V., Vinogradova A. et al. STC-Innovation speaker recognition systems for far-field speaker verification challenge 2020. In Processing INTERSPEECH. 2020, pp. 3466-3470.

8. Sehgal A., Kehtarnavaz N. A convolutional neural network smartphone app for real-time voice activity detection. IEEE Access. 2018, vol. 6, pp. 9017-9026.

9. Sofer A., Chazan S.E. CNN self-attention voice activity detector. arXiv preprint arXiv:2203.02944. 2022.

10. Sohn J., Kim N. S., Sung W. A statistical model-based voice activity detection. IEEE signal processing letters. 1999, vol. 6, no. 1, pp. 1-3.

11. Ramırez J., Segura J.C., Benıtez C., De La Torre A., Rubio A. Efficient voice activity detection algorithms using long-term speech information. Speech communication. 2004, vol. 42, no. 3-4, pp. 271-287.

12. Moattar M.H., Homayounpour M.M. A simple but efficient real-time Voice Activity Detection algorithm. 17th European Signal Processing Conference. 2009, pp. 2549-2553.

13. Pavlyshenko B. Using stacking approaches for machine learning models. Second International Conference on Data Stream Mining & Processing (DSMP). IEEE, 2018, pp. 255-258.

14. Breiman L. Random forests. Machine learning. 2001, vol. 45, no. 1, pp. 5-32.

15. Sagi O., Rokach L. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018, vol. 8, no. 4, 1249 p.

16. Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061. 2020.

17. Juba B., Le H.S. Precision-recall versus accuracy and the role of large data sets. Proceedings of the AAAI conference on artificial intelligence. 2019, vol. 33, no. 1, pp. 4039-4048.

18. Refaeilzadeh P., Tang L., Liu H. Cross-validation. Encyclopedia of database systems. 2009, vol. 5, pp. 532-538.

19. Arlot S., Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010, vol. 4, pp. 40-79.

20. Bergstra J., Bengio Y. Random search for hyper-parameter optimization. Journal of machine learning research. 2012, vol. 13, no. 2, pp. 281-305.

21. Stefanidi A., Topnikov A., Tupitsin G., Priorov A. Application of convolutional neural networks for multimodal identification task. Proceedings of 26th Conference of Open Innovations Association FRUCT. 2020, pp. 423-428.

22. Stefanidi A., Topnikov A., Priorov A, Kosterin I. Modification of VGG neural network architecture for unimodal and multimodal biometrics. Proceedings of 18th IEEE East-West Design & Test Symposium (EWDTS). 2020, pp. 1-4.