Разработка и исследование алгоритма для раздельной записи речи нескольких спикеров

С. Мхаммад, С. А. Молодяков

***Аннотация*— Распознавание речи нескольких одновременно говорящих людей стало важной темой в задачах искусственного интеллекта. Разработан и проанализирован алгоритм раздельной записи голосов нескольких спикеров. Алгоритм включает в себя следующие этапы записи голосовых данных: удаление посторонних звуков, удаление тишины, кластеризацию речевых сегментов с соответствующими кластерными метками, применение нейронной сети для записи текста отдельно для каждого спикера. Особенностью рассматриваемого алгоритма является применение конволюционной нейронной сети на этапе очистки голоса от посторонних звуков. Для записи текста применяется модель Whisper. Она не учитывает случай нескольких спикеров, поэтому перед применением модели вводятся дополнительные шаги. На каждом шаге алгоритма анализируются лучшие методы и метрики для их оценки. Определены метрики, как для отдельных этапов, так и для системы в целом. На основе определения оценок метрик проведено исследование и выделены методы, которые дают наилучшие результаты. При очистке голоса лучший результат дает применение сверточной нейронной сети. На этапе удаления тишины предложено использовать метод на основе обнаружения голосовой активности. При кластеризации речевых сегментов возможно использовать LSTM-модель или сиамскую сеть. Разработано программное приложение, обеспечивающее распознавание и раздельную запись текстов спикеров, говорящих на русском, английском и арабском языках.**

***Ключевые слова*— Конволюционная нейронная сеть, Кластеризация, Машинное обучение, Нейронные алгоритмы, Неконтролируемое обучение, Whisper.**

I. ВВЕДЕНИЕ

Обработка звука стала важнейшей областью исследований и разработок, особенно в задачах записи и синтеза речи и музыки. С развитием методов машинного обучения нейронные сети стали мощным инструментом для распознавания речи. Они позволяют существенно расширить набор решаемых задач. Одной из таких задач является задача раздельной записи одновременного разговора нескольких спикеров. Для решения этой задачи требуется разработать соответствующий алгоритм и программное обеспечение. Известные программные системы распознавания речи не решают в полной мере поставленной задачи. Большинство из них возвращает длинный текст без идентификации говорящего. Хотя методы разделения или диаризация спикеров по аудиозаписи известны Особенность большинства алгоритмов работы с аудио заключается в том, что на каждом этапе работы существует множество методов, которые могут быть использованы для решения задачи. Поэтому важным элементом проектирования является этап анализа и выбора методов для реализации системы. Среди известных методов можно выделить методы, связанные с использованием нейронных сетей. Известны нейронные сети распознавания речи, такие как SpeechRecognition, DeepSpeech, Whisper и другие, но они могут использоваться только в качестве одного из элементов в разрабатываемой системе.

Еще одной задачей при проектировании является необходимость оценки качества работы спроектированной системы. Для оценки качества используются соответствующие метрики. Поэтому необходимо выбрать метрики, как для отдельных этапов, так и для оценки системы в целом.

В статье последовательно представлены этапы разработки программной системы, которые включают рассмотрение известных публикаций, рассмотрение базового алгоритма и анализ известных методов обработки живой или записанной речи, выбор оценочных метрик, описание разработанного алгоритма с выделением новых элементов и представление результатов работы созданного приложения.

II. АНАЛИЗ ИЗВЕСТНЫХ РАБОТ

Для решения проблемы диаризации или разделения спикеров при распознавании речи были разработаны различные алгоритмы. В исследовании [1] использована система верификации спикеров на основе d-векторов с текстонезависимой моделью на основе LSTM. При диаризации использовался только один американский английский язык и уровень ошибок диаризации составлял 12,0%.

В работе [2] предложена сквозная модульная система LibriCSS, которая объединяет независимо обученные компоненты разделения речи, диаризации спикеров и распознавания речи (automatic speech recognition ASR). Для оценки работы системы использовался коэффициент ошибок диаризации (diarization error rate DER), а также коэффициент ошибок в словах (Word error rate WER). Коэффициент DER в системе составил 12,7 %. В исследовании [3] используется полностью

Статья получена 23 февраля 2025.
Мхаммад Салма – Студентка магистратуры Санкт-Петербургского политехнического университета Петра Великого (СПБПУ)
email: mhammad.s@edu.spbstu.ru
Молодяков Сергей Александрович – д.т.н., профессор СПБПУ
email: molodyakov_sa@spbstu.ru.

контролируемый подход к диаризации спикера, названный unbounded interleaved-state recurrent neural networks (UIS-RNN). Модуль кластеризации без контроля был заменен на онлайновый генеративный процесс, который естественным образом включает меченые данные для обучения.

Другие исследования посвящены более конкретным деталям. В работе [4] разработана комплексная модель на основе метода гауссовой смеси для текстонезависимой идентификации голоса с использованием предварительно сегментированного аудио. Обнаружено, что элементы комплексной модели спикера, обученные на кадрах, принадлежащих только к одному широкому фонетическому классу, обладают различной способностью к разделению в зависимости от состава фонетического класса. Несмотря на то, что исследование дало хороший метод идентификации спикера, оно ограничено распознаванием спикера в заранее определенном наборе спикеров.

В исследовании [5] использовалось программное обеспечение с открытым исходным кодом под названием LIUM_SpkDiarization, и оно тестировалось с различным количеством акустических признаков. Результат хороший, но тестирование проводилось только на аудиоданных небольшой длительности, поэтому результаты исследования не были восприняты всерьез.

В работе по результатам анализа публикаций [1-5] определен базовый алгоритм раздельной записи нескольких спикеров. Рассмотрим этот алгоритм, проведем анализ лучших методов его реализации.

III. Базовый алгоритм и методы его реализации

A. *Описание алгоритма*

Базовый алгоритм раздельной записи нескольких спикеров представлен на рис.1. Он включает следующие этапы: удаление посторонних звуков и шумов из входного аудио файла, удаление тишины, кластеризация речевых сегментов с соответствующими кластерными метками, распознавание речи каждого спикера.

На первом этапе происходит удаление посторонних звуков и шумов. Для удаления посторонних источников звука в работах [6] предлагается использовать классические методы спектральной или Wavelet фильтрации, а для удаления шума фильтрацию Винера или фильтрацию Калмана. Однако применение нейронных сетей дает более хорошие результаты. Для обучения нейронных сетей может использоваться маркированный набор данных, который должен содержать чистый и зашумленный звук. В результате нейронная сеть после обучения понимает разницу между шумами и речью.

Удаление тишины из аудио и разделение ее на речевые сегменты является важным шагом в распознавании речи для нескольких говорящих, поскольку он обнаруживает речь каждого говорящего, а также время ее возникновения. Это влияет на качество кластеризации для определения того, кто говорит, и преобразование его речи в текст. Существуют различные методы, которые можно использовать, например статистические подходы, которые могут использовать гауссовские распределения для моделирования различия между речью и тишиной. Сравнение речевых сегментов со смоделированным шумом может привести к более точному обнаружению тишины, как показано в [7]. Однако, как показано в работе [8], метод анализа уровня энергии или мощности аудиосигналов дает хороший результат. Сигнал с энергией ниже определенного порога указывает на тишину. Каждый сегмент занимает время между двумя моментами тишины. Известно, что успешно используется алгоритм WebRTC (Web Real-Time Communication), который строится на основе обнаружения голосовой активности (voice activity detection VAD). Для различения речи и неречи в аудиопотоках используется подход на основе машинного обучения для классификации кадров. Он отвечает за обнаружение речи в смешанном аудиосигнале [9].

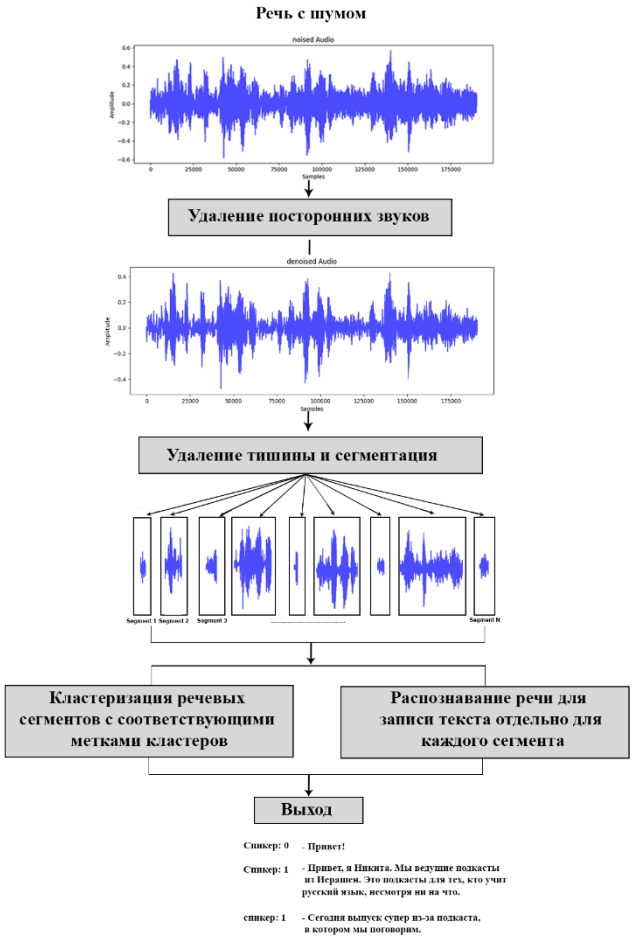

Рис. 1. Базовый алгоритм раздельной записи нескольких спикеров

Следующим этапом является кластеризация речевых сегментов. На этом этапе происходит группировка сегментов, которые принадлежат одному и тому же спикеру без предварительного знания участвующих в разговоре. Здесь используются методы извлечения признаков либо из временной области, либо из частотной области. Основным методом является использование мел-кепстральных коэффициентов (Mel-

Frequency Cepstral MFCC) [10].

Для кластеризации можно использовать различные алгоритмы, такие, как K-средние, модели гауссовой смеси (Gaussian mixture models GMM) и агломеративная иерархическая кластеризация. В работе [11] показано, что использование k-средних дает хороший результат. В работе [12] показано, что использование метода GMM позволяет классифицировать эмоций из речи.

Можно использовать методы кластеризации, связанные с разными моделями нейронных сетей, такими как LSTM (Long short-term memory), сеть с самоорганизующейся картой (Self-Organizing Maps, SOM), сиамская сеть (Siamese Network, SN) [13-15]. При использовании сетей типа LSTM происходит сжатие входных данных в представление с меньшей размерностью (кодер), а затем реконструирование выходных данных из этого представления (декодер). Выходной слой решает задачу кластеризации [13]. Архитектура сети SOM содержит в дополнение к входным и выходным слоям слой карты, который обычно состоит из сетки нейронов, расположенных в двумерном пространстве. Каждый нейрон представляет кластер схожих входных шаблонов [14].

Распознавание речи в текст в настоящее время построено на применении нейронных сетей. Одной из самых мощных нейронных сетей, имеющей архитектуру Transformer и открытый исходный код является Whisper. Она применяется для распознавания большого количества современных языков, в том числе арабского. Хотя похожие результаты при распознавании русского и других языков можно получить и при применении сетей DeepSpeech, Conformer, SpeechRecognition.

На выходе работы алгоритма получается распределенный по спикерам текст. Для каждого аудио сегмента получается своя строка текста.

IV. Разработанный алгоритм и его реализация

Разработан алгоритм для раздельной записи речи нескольких спикеров (рис.2). Он отличается использованием нейронной сети для очистки аудио (denoising). При исследовании алгоритма проведена оценка возможности применения нескольких методов кластеризации. Применялись следующие методы: K-means, GMM, LSTMs, SOMs, SN.

A. Применение конволюционной нейронной сети для очистки аудио

Для очистки аудиосигнала выбрана модель сверточной нейронной сети (convolutional neural network, CNN). Значительным преимуществом такой сети является высокая скорость процессов обучения и вывода. Скорость достигается за параллельного применения нескольких фильтров, а как следствие использование возможностей графических процессоров (GPU). Для обучения сети используется набор данных VoiceBank + DEMAND, в котором сочетаются чистые образцы речи из корпуса Voice Bank и различные фоновые шумы, взятые из базы данных DEMAND. Он содержит чистый набор данных и набор шумов. Набор данных разделили на обучающий, проверочный и тестовый, например: 80% данных для обучения, 10% для проверки и 10% для тестирования.

После завершения обучения веса модели сохранили в файл, чтобы повторно использовать обученную модель для тестирования. Затем рассчитали метрики выбора для оценки модели и проверили, достигает ли результат принятого диапазона. Как видно на графике 1 (рис.1), в зашумленном аудио практически нет различий между сегментами речи и тишины, что влияет на процесс сегментации и увеличивает ошибку в результатах распознавания речи. Но после использования модели

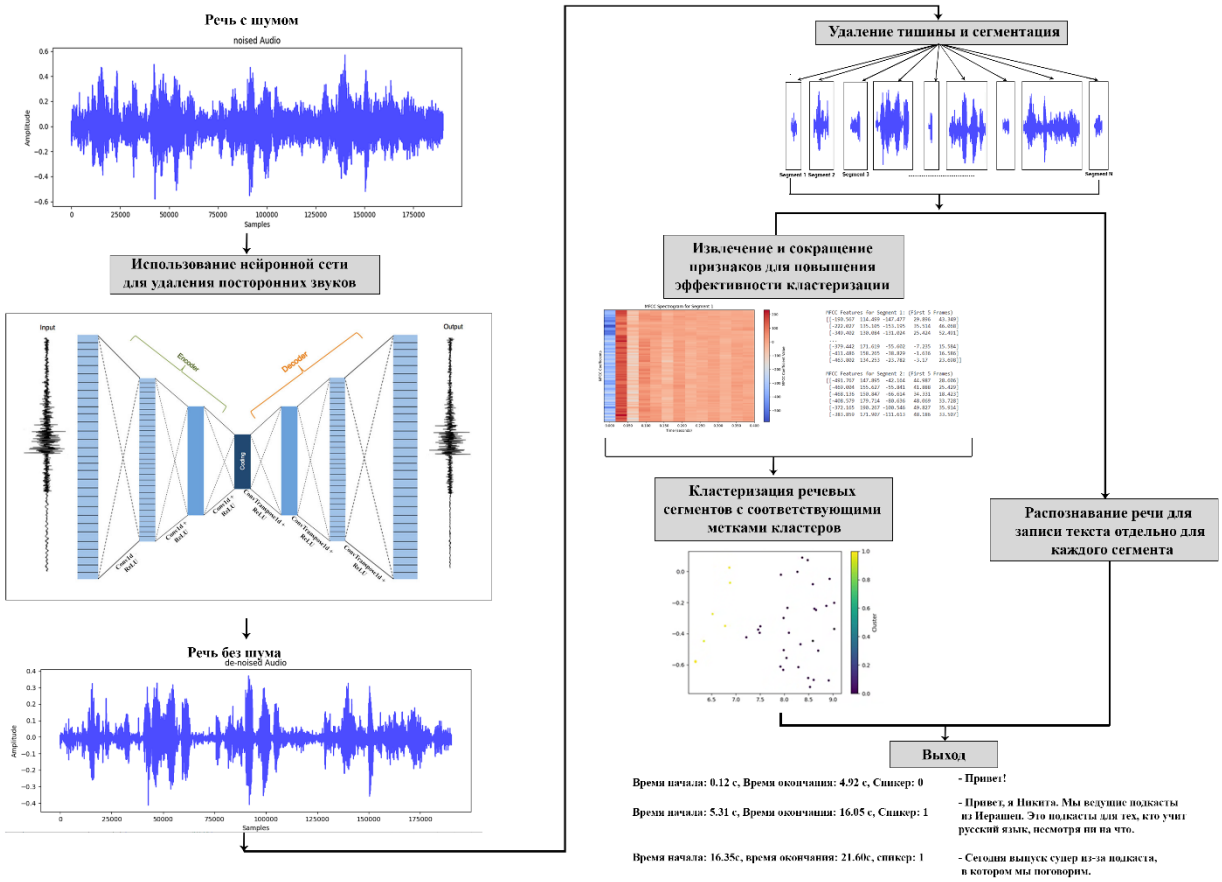

Рис. 2. Разработанный алгоритм для раздельной записи речи нескольких спикеров

удаления шума на графике 2 (рис.2) сегмент речи стал более четким для сегментации с помощью рассмотренных методов.

B. Удаление тишины и сегментирования речи

Для удаления тишины использовались два метода: на основе порога энергии аудиосигнала и на основе обнаружения голосовой активности VAD. В первом методе энергия аудиосигнала рассчитывалась в соответствии с формулой (1), где n – номер отсчета, N - общее количество отсчетов в речевом кадре, k - номер кадра [16]:

$$E_k = \sum_{n=1}^{N} s_k^2(n) \qquad (1)$$

Задается порог, затем он сравнивается с уровнем аудиосигнала, выявляются кадры, в которых амплитуда сигнала ниже порога. Эти кадры помечаются как тишина. В результате происходит разделение аудио на сегменты с указанием времени начала и окончания каждого из них [17].

При обнаружении голосовой активности VAD используются более сложные методы. Сначала сигнал делится на короткие перекрывающиеся кадры, чтобы проанализировать изменения во времени. Затем, как и в предыдущем описании, вычисляется порог для определения тихих и голосовых кадров. Вокальные кадры могут содержать речевые и неречевые звуки. Модели машинного обучения, такие как CNN и RNN, используются для классификации аудиокадров как речевых или неречевых, в зависимости от случайности распределения частот (спектральной энтропии) в аудиокадрах [18]. В разработанном приложении использована модель CNN, которая обучена на различных наборах данных для повышения точности. Для создания модели использовалась функция *webrtcvad.Vad()*, а для определения наличия в кадре речевых и неречевых звуков применялась функция *vad.is_speech()*. (табл. 1).

C. Извлечение признаков из сегментов

Для извлечения признаков применялся метод мел-кепстральных коэффициентов (MFCC). Для вычисления MFCC-коэффициентов использовались следующие шаги. Каждый кадр аудиосигнала подвергается оконной обработке, чтобы минимизировать краевые эффекты при вычислениях преобразования Фурье. После этого применяется преобразование Фурье для получения частотного спектра для каждого кадра. Затем частотные отсчеты умножаются на треугольные полосовые фильтры [19]. Коэффициенты MFCC представляют собой числовые представления спектральной формы звука. Для вычисления MFCC-коэффициентов использовались функции *feature.mfcc()* и *resample()* из библиотеки *librosa* (табл.1).

D. Кластеризация речевых сегментов

Для кластеризации сегментов в отдельные группы спикеров на основе коэффициентов MFCC проведено сравнение следующих методов: K-means, GMM, LSTMs, SOMs, SN. Цель кластеризации - минимизировать дисперсию внутри каждого кластера и максимизировать дисперсию между различными кластерами. Алгоритм K- средних объединяет обучение без контроля с векторным квантованием [20]. Для кластеризации данных на k групп, использовалась функция *KMeans()* из библиотеки sklearn.cluster (табл.1). Для реализации модели GMM была использована функция *GaussianMixture()* из библиотеки *sklearn.mixture* (табл. 1). В обеих библиотеках есть функция *fit_predict()*, которая также использовалась для подгонки модели к данным и возврата кластерных меток для каждого образца.

Для определения модели LSTM использована функция *LSTM()* из библиотеки *lstm_model*, а функция *predict()* - для генерации предсказаний (кластерных меток) из модели LSTM. Для нейронной сети с SOM, использована библиотека *minisom* с функцией *MiniSom()* для ее инициализации. Для модели SN использованы функции *Conv1d()* и *Dense()* из библиотеки *tensorflow* для создания конкретной модели.

E. Распознавание речи и разделение по спикерам

Для распознавания речи использована модель машинного обучения Whisper. Она была обучена на 680 000 часов многоязычных данных, собранных из Интернета. [21] Whisper может быть использован для таких задач, как идентификация языка, временные метки на уровне фраз и транскрипция речи на нескольких языках, например английском, русском и арабском. Основные компоненты этой архитектуры включают: кодер, который отвечает за преобразование входного аудиосигнала в представление признаков, и декодер, который предсказывает вывод текста на основе закодированных представлений.

Работа алгоритма опробована на аудиозаписях с разных языков. Видно, что он хорошо работает при генерации диалогов на арабском, русском и английском языках.

Start Time: 0.66s, End Time: 7.83s, Speaker: 1, Transcription: صباح الخير صباح النور أهلا بك
Start Time: 9.42s, End Time: 12.75s, Speaker: 0, Transcription: أهلا ومرحبا كيف حالك
Start Time: 14.34s, End Time: 16.41s, Speaker: 0, Transcription: بخير والحمد لله
Start Time: 16.86s, End Time: 18.48s, Speaker: 0, Transcription: وكيف حالك أنت
Start Time: 19.47s, End Time: 21.27s, Speaker: 1, Transcription: بخير والحمد لله
Start Time: 21.81s, End Time: 25.41s, Speaker: 1, Transcription: هل انتطالب جديد في هذه المدرسة؟

Start Time: 0.12s, End Time: 4.92s, Speaker: 0, Transcription: Привет!
Start Time: 5.31s, End Time: 16.05s, Speaker: 1, Transcription: Привет, я Никита. Мы ведущие подкасты из Иерашен. Это подкасты для тех, кто учит русский язык, несмотря ни на что.
Start Time: 16.35s, End Time: 21.60s, Speaker: 1, Transcription: Сегодня выпуск супер из-за подкаста, в котором мы поговорим.
Start Time: 22.23s, End Time: 23.10s, Speaker: 0, Transcription: Просмотрим в токе.

Start Time: 0.00s, End Time: 9.36s, Speaker: 0, Transcription: I said, well, you know, that part of it from home, any familiar face is a comfort. Even if it is just my price. Yeah. Yeah. Yeah.
Start Time: 10.32s, End Time: 23.07s, Speaker: 1, Transcription: Anyway, that was fun, the food was great, it was catered, it was just wonderful. Wow. And I won't tell you everything because it's all in your letter. Okay, yeah, I notice that every time I get a letter, I said, well we've covered all this.
Start Time: 23.43s, End Time: 24.96s, Speaker: 0, Transcription: And uh...

F. Описание программного обеспечения

Программное обеспечение разработано на языке python. Для модели удаления шума использовалась библиотека torch с функциями *Conv1d()* для определения кодера и *ConvTranspose1d()* для определения декодера. Определенные слои были сложены для архитектуры модели с помощью функции *Sequential()*, модель была обучена на наборе данных VoiceBank+DEMAND. Во время обучения использовали функцию *optim.Adam()* для оптимизации параметров модели.

Для удаления тишины были определены два метода. Первый вычисляет энергию и порог с помощью библиотеки *numpy* для разделения аудио на сегменты. Второй использует библиотеку *webrtcvad* с функцией *vad.is_speech()*. Для выбора лучшего метода использованы математические функции из библиотеки *numpy* и также библиотеки *sklearn.metrics* для расчета метрик DER и Precision.

Многие функции из библиотеки sklearn.metrics используются для оценки предложенных методов кластеризации (K-means, GMM, LSTMs, SOMs, SN) и расчета метрик в зависимости от набора данных CallHome. Данный набор содержит аудиофайлы для двух спикеров с информацией о сегментах, таких как время начала, время окончания, и кто говорит. Эта информация важна для оценки эффективности разработанного алгоритма. Библиотека *Whisper* использовалась для генерации текста из аудиосегментов с помощью функции *transcribe()*. В табл. 1 приведены использованные функции с их описанием.

TABLE I. БИБЛИОТЕКИ И ФУНКЦИИ С ОПИСАНИЯМИ

Библиотека	***Функция***	***Описание***
torch	*optim.Adam()*	реализовать алгоритм оптимизации Адама для обучения модели.
	nn.Sequential()	Последовательный контейнер для укладки слоев для архитектуры модели.
	nn.Conv1d()	Конволюционный слой 1D, используемый для извлечения признаков из аудиосигналов.
	nn.ConvTranspose 1d()	1D транспонированный конволюционный слой, используемый для апсемплинга.
librosa	*resample()*	Передискретизировать аудиосигнал до заданной частоты дискретизации.
	feature.mfcc()	Вычислить MFCC из аудиосигнала.
webrtcvad	*webrtcvad.Vad()*	Инициализировать объект VAD.
	vad.is_speech()	Определить, содержит ли данный аудиокадр речь.
whisper	*whisper_model.tra nscribe()*	Транскрибировать аудио в текст с помощью модели Whisper.
sklearn	*KMeans()*	Применить алгоритм кластеризации K-Means для группировки точек данных.
	silhouette_score()	Оценить качество кластера с помощью коэффициента силуэта.
lstm_model	*LSTM()*	Добавить слой LSTM (Long Short-Term Memory) для улучшения входных данных кластеризации.
	predict()	Генерировать предсказания (эмбеддинги) из LSTM-модели.
minisom	*MiniSom()*	Инициализировать самоорганизующуюся карту (SOM) для улучшения входных данных кластеризации.
	train_random()	Обучить SOM, используя случайно выбранные образцы из набора данных.

V. МЕТРИКИ И ОЦЕНКА РАЗРАБОТАННОЙ СИСТЕМЫ

Для оценки эффективности разработанной системы важно выбрать метрики, которые позволили бы сравнивать разработанную систему с другими. Были выбраны метрики для каждого из этапов работы алгоритма.

A. Метрики удаления шумов

Средняя квадратичная ошибка (MSE) определяет среднюю квадратичную разницу между исходным чистым аудио и деноизированным аудио, полученным с помощью модели (2). Она дает количественную оценку точности реконструкции [22].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 , \qquad (2)$$

где:

- y_i - исходный аудиоотсчет,
- $\hat{y}_i$ - деноизированный аудиоотсчет,
- N - общее количество отсчетов

По значению MSE можно рассчитать пиковое отношение сигнал/шум (PSNR), которое представляет собой логарифмическую меру соотношения между максимально возможной мощностью сигнала и мощностью искажающего шума, влияющего на его представление, как показано в (3). Более высокое значение PSNR указывает на лучшее качество реконструкции.

$$PSN = 10\log\left(\frac{MAX^2}{MSE}\right) \qquad (3)$$

В табл. 2 показан результат работы разработанной системы после завершения обучения модели с идеальными значениями. Видно, что очистка от шумов аудио работает эффективно.

TABLE II. МЕТРИКИ УДАЛЕНИЯ ШУМА

Метрика	***Требуемый результат***	***Разработанная система***
MSE	НИЖЕ 0,01 (НОРМАЛИЗОВАННЫЙ ЗВУК)	0.0004
PSNR	30+ DB (ВЫСОКОЕ КАЧЕСТВО)	34.86

B. Метрики удаления тишины и сегментации

Коэффициент ошибок диаризации DER - это метрика, используемая для оценки точности систем дикторской диаризации, которые предназначены для идентификации и сегментации отдельных спикеров в аудиопотоке. Для ее выполнения используется набор данных аудиозаписей нескольких спикеров, содержащих информацию «кто когда говорил», для тестирования методов сегментации речи и выбора наилучшего варианта. Расчет DER (4) включает в себя три основных типа ошибок [23]:

- Ложный спикер (False Speaker): Продолжительность неречевых высказываний, которые были ошибочно классифицированы как речь.
- Пропущенное обнаружение (Missed Detections): Продолжительность речи, которая была ошибочно классифицирована как неречевая.
- Путаница с спикером (Speaker Confusion): Продолжительность речи, которая ошибочно идентифицируется как принадлежащая другому диктору.

$$\text{DER} = \frac{\text{False Speaker+ Missed Detections+ Speaker Confusion}}{\text{Total Ground Truth Speech Duration}} \qquad (4)$$

Идеальное значение коэффициента ошибок диаризации равно 0. Однако достижение значения DER, равного 0, обычно нереально. Таким образом, хорошие значения могут варьироваться в зависимости от сложности задачи и возможностей системы, но в целом приемлемый коэффициент ошибок обычно определяется

как DER ≤ 15 %, а отличнй считается как DER ≤ 5 %.

Модули обнаружения изменений в сегментации можно оценивать с помощью метрик в точность Precision (5) и чувствительность Recall (6). Точность определяет точность положительных предсказаний, сделанных моделью. Она отвечает на вопрос, сколько из обнаруженных точек изменений были действительно правильными [24].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \qquad (5)$$

где True Positives (TP) - количество правильно обнаруженных изменений, а False Positives (FP) - количество сегментов, неверно идентифицированных как изменения. Recall определяет способность системы обнаруживать все реальные точки изменений. Ее можно представить в виде (6), где False Negatives (FN) - это количество фактических изменений, которые были пропущены моделью. Превосходное значение точности и Recall составляет более 80%, а хорошее значение - около 60%.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (6)$$

Для оценки работы алгоритма на этапах удаления тишины и сегментации речи проведено сравнение методов: основанного на пороговых значениях, с предварительно подготовленной моделью VAD. Результаты приведенные в табл. 3.

TABLE III. СРАВНЕНИЕ ПРЕДЛОЖЕННЫХ МЕТОДОВ

	Метрика	***DER***	***Precision***	***Recall***
Без модели удаления шума	***Threshold -Based***	0.115	0.46	0.51
	VAD	0.1126	0.52	0.53
С помощью модели удаления шума	***Threshold -Based***	0.1061	0.51	0.54
	VAD	0.102	0.52	0.54

Из табл. 3 видно, что в зависимости от оценки DER использование VAD дает лучший результат, чем использование метода, основанного на пороговых значениях. Кроме того, показатели Precision и Recall дают лучшие результаты при использовании VAD. Значение DER стало лучше, когда мы использовали те же методы с применением модели удаления шума. Это позволяет сделать вывод о том, что использование данной модели было полезно для улучшения работы рассматриваемого алгоритма.

За счет использования нейронной сети для удаления шума параметр DER был снижен до 10,2%, что лучше, чем приводятся в работах [1, 3]. Однако деноизированный звук все равно содержит некоторые шумы, которые модель не смогла удалить. Эту проблему можно решить, используя набор данных с гораздо более широким и большим количеством шумов.

C. Метрики кластеризации

Silhouette Score - одна из самых популярных метрик для кластеризации, она определяет, насколько объект похож на свой собственный кластер по сравнению с другими кластерами. Значение варьируется от -1 до 1, где более высокое значение указывает на то, что точка данных хорошо соответствует своему кластеру и плохо - соседним кластерам. Однако хорошим значением Silhouette Score считается значение выше 0,5, что свидетельствует о хорошей кластеризации [25]. Метрику Silhouette Score можно определить в соответствии с (7).

$$\text{Silhouette Score} = \frac{b_i - a_i}{\max(a_i, b_i)}, \qquad (7)$$

где a_i - среднее расстояние от каждой точки до других точек данных в пределах одного кластера, а b_i - среднее расстояние от каждой точки до всех других кластеров, в которые она не входит, вычисляется так.

Другой важный метод, который следует принимать во внимание, - Calinski-Harabasz Index. Этот индекс оценивает кластеризацию, сравнивая дисперсию между кластерами с дисперсией внутри кластеров. Более высокие значения указывают на лучшую конфигурацию кластеризации. Индекс можно рассчитать по формуле (8), где BSS - межкластерная сумма квадратов, а WSS - внутрикластерная сумма квадратов при n точках данных.

$$\text{CalinskiHarabasz Index} = \frac{BSS/(k-1)}{WSS/(n-k)} \qquad (8)$$

Accuracy Score также является хорошей метрикой кластеризации, которая измеряет количество правильных предсказаний в задачах классификации. Его значение должно находиться в диапазоне от 0 до 1, и чем оно выше, тем лучше [26]. Рассчитать показатель точности можно по следующей формуле:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{Total\ \text{Predictions}} \qquad (9)$$

Для того чтобы определить наилучший метод кластеризации, проведено тестирование методов в разработанном приложении. Получены соответствующие метрики. Результаты представлены в табл. 4.

TABLE IV. СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ КЛАСТЕРИЗАЦИИ

	Threshold-Based			***VAD***		
Метрика	***Silhouette Score***	***Calinski-Harabasz Index***	***Accuracy***	***Silhouette Score***	***Calinski-Harabasz Index***	***Accuracy***
K-means	0.5302	60.3019	0.5762	0.6354	72.340	0.5263
GMM	0.4218	28.0301	0.5423	0.63549	72.3406	0.5263
LSTMs	0.6994	241.499	0.55	0.7230	158.048	0.6052
SOMs	0.39052	43.254	0.57627	0.3285	18.5615	0.5263
SN	0.71421	156.296	0.5737	0.68813	99.0866	0.486

Из табл. 4 видно, что измеренные значения метрик кластеризации меняются в зависимости от того, какие методы используются для сегментации аудио и создания аудиогрупп для спикеров в первый раз. При использовании традиционного порогового метода и

метода k-means для кластеризации мы получили приемлемый результат по всем трем метрикам. Но при использовании того же метода с добавлением LSTM-модели результат оказался гораздо лучше. То же самое можно сказать и об использовании VAD. В сравнении с этим, использование SOM с контролируемыми данными дает наихудший результат при обоих способах сегментации. Наконец, мы увидели, что хороший результат был достигнут при добавлении как модели LSTM с VAD, так и сиамских сетей с пороговым методом, которые считаются лучшими вариантами.

VI. ВЫВОДЫ

В работе представлен алгоритм для раздельной записи речи нескольких спикеров. На основе рассмотрения основных шагов алгоритма проведен анализ лучших методов для каждого шага. Определено, что выбор метода сегментирования речи влияет на результат процесса обнаружения и идентификации спикеров из-за его влияния на группу кластеров. Более того, обнаружено, что наиболее эффективным способом реализации является использование VAD для удаления тишины и сегментирования речи вместе с LSTM-сетями для кластеризации и определения спикеров.

В соответствии с полученными результатами разработанный алгоритм как минимум не хуже алгоритмов, описанных в известных работах. В этом исследовании не упоминается в деталях и не тестируется с помощью оценочных метрик модель Whisper. Она известна и описана в работах, как мощный инструмент преобразования речи в текст с различных языков.

БИБЛИОГРАФИЯ

[1] Wang Q. et al. Speaker diarization with LSTM //2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2018. – С. 5239-5243.

[2] Raj D. et al. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis //2021 IEEE spoken language technology workshop (SLT). – IEEE, 2021. – С. 897-904.

[3] Zhang A. et al. Fully supervised speaker diarization //ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2019. – С. 6301-6305.

[4] Ермоленко Т. В., Клименко Н. С. Использование сегментации речевого сигнала для построения комплексной модели диктора в системе идентификации говорящего //Информатика и автоматизация. – 2013. – Т. 3. – №. 26. – С. 332-348.

[5] Рогов А. А., Петров Е. А. Анализ существующих свободно распространяемых систем разделения дикторов на фонограмме //Фундаментальные исследования. – 2015. – №. 6-1. – С. 67-72.

[6] Ласточкин А. В., Кобелев В. Ю. Метод удаления шума на основе вейвлет-обработки, адаптированный к разрывным сигналам: тр. 5-й Междунар. конф. “Цифровая обработка сигналов и её применение (DSPA-2003)” [Электронный ресурс] // С.-Пб.: ЗАО АВТЭКС, 2003. – Режим доступа: http://www.autex.spb.ru (дата обращения: ноябрь 2024).

[7] Sahoo T. R., Patra S. Silence removal and endpoint detection of speech signal for text independent speaker identification //International Journal of Image, Graphics and Signal Processing. – 2014. – Т. 6. – №. 6. – С. 27.

[8] Hanifa R. M. et al. Voiced and unvoiced separation in Malay speech using zero crossing rate and energy //Indones. J. Electr. Eng. Comput. Sci. – 2019. – Т. 16. – №. 2. – С. 775-780.

[9] Ball J. Voice Activity Detection (VAD) in Noisy Environments //arXiv preprint arXiv:2312.05815. – 2023.

[10] Abdul Z. K., Al-Talabani A. K. Mel frequency cepstral coefficient and its applications: A review //IEEE Access. – 2022. – Т. 10. – С. 122136-122158.

[11] Abdul Z. K., Al-Talabani A. K. Mel frequency cepstral coefficient and its applications: A review //IEEE Access. – 2022. – Т. 10. – С. 122136-122158.

[12] Bhukya R. K., Raj A. Automatic speaker verification spoof detection and countermeasures using gaussian mixture model //2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). – IEEE, 2022. – С. 1-6.

[13] Sinaga K. P., Yang M. S. Unsupervised K-means clustering algorithm //IEEE access. – 2020. – Т. 8. – С. 80716-80727.

[14] An S., Ling Z., Dai L. Emotional statistical parametric speech synthesis using LSTM-RNNs //2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). – IEEE, 2017. – С. 1613-1616.

[15] Analytics B. et al. Self-Organizing Map and Multidimensional Scaling in a Tandem Approach: a Visualization of Bankruptcy Trajectory. – 2019.

[16] Yeo J. H. et al. Visual speech recognition for low-resource languages with automatic labels from Whisper model // arXiv preprint arXiv:2309.08535. – 2023.

[17] Warule P., Mishra S. P., Deb S. Significance of voiced and unvoiced speech segments for the detection of common cold //Signal, image and video processing. – 2023. – Т. 17. – №. 5. – С. 1785-1792.

[18] Andersen L. R., Jacobsen L. J., Campos D. Compressed, Real-Time Voice Activity Detection with Open Source Implementation for Small Devices //Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence. – 2023. – С. 1-10.

[19] Ashar A., Bhatti M. S., Mushtaq U. Speaker identification using a hybrid cnn-mfcc approach //2020 International Conference on Emerging Trends in Smart Technologies (ICETST). – IEEE, 2020. – С. 1-4.

[20] Liu J. et al. A hybrid news recommendation algorithm based on K-means clustering and collaborative filtering //Journal of Physics: Conference Series. – IOP Publishing, 2021. – Т. 1881. – №. 3. – С. 032050.

[21] Oruh J., Viriri S., Adegun A. Long short-term memory recurrent neural network for automatic speech recognition // IEEE Access. – 2022. – Т. 10. – С. 30069–30079.

[22] Khan U., Hernando Pericás F. J. Unsupervised training of siamese networks for speaker verification //Interspeech 2020: the 20th Annual Conference of the International Speech Communication Association: 25-29 October 2020: Shanghai, China. – International Speech Communication Association (ISCA), 2020. – С. 3002-3006.

[23] Kang H., Park C., Yang H. Evaluation of Denoising Performance of ResNet Deep Learning Model for Ultrasound Images Corresponding to Two Frequency Parameters // Bioengineering. – 2024. – Т. 11, № 7.

[24] Arora M., Kanjilal U., Varshney D. Evaluation of information retrieval: precision and recall //International Journal of Indian Culture and Business Management. – 2016. – Т. 12. – №. 2. – С. 224-236.

[25] Punhani A. et al. Binning-based silhouette approach to find the optimal cluster using K-means //IEEE Access. – 2022. – Т. 10. – С. 115025-115032.

[26] Suraya S., Sholeh M., Lestari U. Evaluation of Data Clustering Accuracy using K-Means Algorithm //International Journal of Multidisciplinary Approach Research and Science. – 2023. – Т. 2. – №. 01. – С. 385-396.

Developing and analyzing an algorithm for separate speech recording of multiple speakers

S. Mhammad, S. A. Molodyakov

***Abstract*— Speech recognition of multiple simultaneous speakers has become an important topic in artificial intelligence tasks. An algorithm for separate voice recording of multiple speakers is developed and analyzed. The algorithm includes the following stages: removal of extraneous sounds, removal of silence, clustering of speech segments with corresponding cluster labels, application of neural network to record text separately for each speaker. The particular feature of the considered algorithm is the application of convolutional neural network at the stage of voice cleaning from extraneous sounds. The Whisper model is used for text recording. It does not take into account the case of multiple speakers, so additional steps are introduced before applying the model. In each step of the algorithm, the best methods and metrics are analyzed. Metrics are defined for both individual steps and the system as a whole. Based on the determination of the metrics evaluations, a study is done and the methods that give the best results are highlighted. In the voice cleaning stage, the best result is given by the application of convolutional neural network. In silence removal stage, a method based on voice activity detection is proposed. When clustering speech segments, it is possible to use LSTM model or Siamese network. A software application has been developed to recognize and separately record the texts of speakers with Russian, English and Arabic language.**

***Keywords*— Convolutional neural network, Clustering, Machine learning, Neural algorithms, Unsupervised learning, Whisper.**

REFERENCES

[1] Wang Q. et al. Speaker diarization with LSTM //2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2018. – С. 5239-5243.

[2] Raj D. et al. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis //2021 IEEE spoken language technology workshop (SLT). – IEEE, 2021. – С. 897-904.

[3] Zhang A. et al. Fully supervised speaker diarization //ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2019. – С. 6301-6305.

[4] Ermolenko T. V., Klimenko N. S. Using speech signal segmentation to build a complex speaker model in the speaker identification system // Informatics and Automation. - 2013. - Т. 3. - №. 26. - С. 332-348.

[5] Rogov A. A., Petrov E. A. Analysis of existing freely distributed systems of speaker separation on phonogram // Fundamental Research. - 2015. - №. 6-1. - С. 67-72.

[6] Lastochkin A. V., Kobelev V. Yu. V. V., Kobelev V. Yu. A method of noise removal based on wavelet processing adapted to discontinuous signals: Proc. of the 5th Intern. 5th International Conf. “Digital Signal Processing and its Application (DSPA-2003)” [in Russian]. [Electronic resource] // St.-Petersburg: ZAO AVTEKS, 2003. - Access mode: http://www.autex.spb.ru (access date: November 2024).

[7] Sahoo T. R., Patra S. Silence removal and endpoint detection of speech signal for text independent speaker identification //International Journal of Image, Graphics and Signal Processing. – 2014. – Т. 6. – №. 6. – С. 27.

[8] Hanifa R. M. et al. Voiced and unvoiced separation in Malay speech using zero crossing rate and energy //Indones. J. Electr. Eng. Comput. Sci. – 2019. – Т. 16. – №. 2. – С. 775-780.

[9] Ball J. Voice Activity Detection (VAD) in Noisy Environments //arXiv preprint arXiv:2312.05815. – 2023.

[10] Abdul Z. K., Al-Talabani A. K. Mel frequency cepstral coefficient and its applications: A review //IEEE Access. – 2022. – Т. 10. – С. 122136-122158.

[11] Abdul Z. K., Al-Talabani A. K. Mel frequency cepstral coefficient and its applications: A review //IEEE Access. – 2022. – Т. 10. – С. 122136-122158.

[12] Bhukya R. K., Raj A. Automatic speaker verification spoof detection and countermeasures using gaussian mixture model //2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). – IEEE, 2022. – С. 1-6.

[13] Sinaga K. P., Yang M. S. Unsupervised K-means clustering algorithm //IEEE access. – 2020. – Т. 8. – С. 80716-80727.

[14] An S., Ling Z., Dai L. Emotional statistical parametric speech synthesis using LSTM-RNNs //2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). – IEEE, 2017. – С. 1613-1616.

[15] Analytics B. et al. Self-Organizing Map and Multidimensional Scaling in a Tandem Approach: a Visualization of Bankruptcy Trajectory. – 2019.

[16] Yeo J. H. et al. Visual speech recognition for low-resource languages with automatic labels from Whisper model // arXiv preprint arXiv:2309.08535. – 2023.

[17] Warule P., Mishra S. P., Deb S. Significance of voiced and unvoiced speech segments for the detection of common cold //Signal, image and video processing. – 2023. – Т. 17. – №. 5. – С. 1785-1792.

[18] Andersen L. R., Jacobsen L. J., Campos D. Compressed, Real-Time Voice Activity Detection with Open Source Implementation for Small Devices //Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence. – 2023. – С. 1-10.

[19] Ashar A., Bhatti M. S., Mushtaq U. Speaker identification using a hybrid cnn-mfcc approach //2020 International Conference on Emerging Trends in Smart Technologies (ICETST). – IEEE, 2020. – С. 1-4.

[20] Liu J. et al. A hybrid news recommendation algorithm based on K-means clustering and collaborative filtering //Journal of Physics: Conference Series. – IOP Publishing, 2021. – Т. 1881. – №. 3. – С. 032050.

[21] Oruh J., Viriri S., Adegun A. Long short-term memory recurrent neural network for automatic speech recognition // IEEE Access. – 2022. – Т. 10. – С. 30069–30079.

[22] Khan U., Hernando Pericás F. J. Unsupervised training of siamese networks for speaker verification //Interspeech 2020: the 20th Annual Conference of the International Speech Communication Association: 25-29 October 2020: Shanghai, China. – International Speech Communication Association (ISCA), 2020. – С. 3002-3006.

[23] Kang H., Park C., Yang H. Evaluation of Denoising Performance of ResNet Deep Learning Model for Ultrasound Images Corresponding to Two Frequency Parameters // Bioengineering. – 2024. – Т. 11, № 7.

[24] Arora M., Kanjilal U., Varshney D. Evaluation of information retrieval: precision and recall //International Journal of Indian Culture and Business Management. – 2016. – Т. 12. – №. 2. – С. 224-236.

[25] Punhani A. et al. Binning-based silhouette approach to find the optimal cluster using K-means //IEEE Access. – 2022. – Т. 10. – С. 115025-115032.

[26] Suraya S., Sholeh M., Lestari U. Evaluation of Data Clustering Accuracy using K-Means Algorithm //International Journal of Multidisciplinary Approach Research and Science. – 2023. – Т. 2. – №. 01. – С. 385-396.