Интеграционный подход распознавания зашумленной русскоязычной речи

Даниил Гомонюк*, Игорь Никифоров†, Дмитрий Дробинцев‡
Высшая школа программной инженерии
Санкт-Петербургский Политехнический Университет
Санкт-Петербург, Россия
Email: *dan.gomonuk@gmail.com, †igor.nikiforovv@gmail.com, ‡drobintsev_df@spbstu.ru

***Аннотация*—Исследовательская работа посвящена методам автоматического преобразования аудиозаписей в текстовый формат - распознаванию речи.**

В частности, особое внимание уделено распознаванию зашумленной русской речи.

В работе предоставления обзор существующих методов распознавания, которые включают "интегральные"и "гибридные"методы. Приведен сравнительный обзор существующих реализаций рассмотренных методов и их метрики. На основе сравнительного анализа делается вывод, что технология "Mozilla DeepSpeech"наиболее мощный инструмент распознавания.

Отличительной особенностью работы является использование комбинированного метода распознавания, который позволяет улучшить качество распознавания зашумленной речи. Комбинированный метод объединяет в себе "интегральные"и "гибридные"методы. Предлагаемый подход реализован в программном средстве для распознавания зашумленной русской речи с использованием технологии "Mozilla DeepSpeech". Результаты показывают эффективность предложенного подхода.

Разработанное программное средство может быть использовано компаниям в целях снижения трудозатрат при осуществлении технической поддержки заказчиков.

***Ключевые понятия*—распознавание речи, зашумленная речь, аудиозапись, Mozilla DeepSpeech, Baidu, Kaldi**

I. Введение

Инновационные подходы и технологии с каждым днем все больше и больше интегрируются в устоявшихся годами сферах жизнедеятельности человека. Не является исключением и применение методов машинного обучения для распознавания аудиозаписей. Так, например, распознавание речи по аудиозаписям позволяет повысить эффективность служб клиентской поддержки, даёт возможность проводить аналитику звонков [1], избегая проблем с соблюдением закона “О персональных данных”, так как зачастую в аудио-звонках упоминается конфиденциальная информация [2]. Ниже перечислены две основные группы методов: основанные на применении скрытых марковских моделей и методы, основанные на нейронных сетях.

Описание методов, основанных на применении скрытых марковских моделей (далее СММ), можно найти, например, в работе [3]. Инструменты на основе этих методов очень точны, но требуют составления словаря, соотносящего слово и его фонемы (например слово “ноль” разбивается на фонемы “n” “oo” “ll”). Такая система не сможет распознать слово которого нет в словаре, но чем больше количество слов, входящих в словарь, тем больше вероятность ошибки, так как выбор слова из словаря становиться неоднозначнее. Такие инструменты подходят для распознавания заранее известных фраз, например речевых команд, но они не эффективны при распознавании спонтанной речи. Одной из систем, реализующих рассматриваемый метод, является CMUSphinx.

Методы, основанные на нейронных сетях, можно разделить на "интегральные"и "гибридные"[4] методы. Несмотря на то, что они подходят для распознавания спонтанной речи, они не избавлены от недостатков:

- качество распознавания во многом зависит от качества исходной аудиозаписи, что накладывает высокие требования на качество исходной аудиозаписи;
- отсутствие универсальных методов и реализующих их библиотек. Зачастую для каждой конкретной задачи необходимо создавать свое собственное решение;
- для каждого языка (например, русского, английского, китайского), приходится проводить дополнительную настройку систем распознавания.

Поэтому актуальной является задача создания такого метода, который бы позволял снизить влияние перечисленных недостатков и повысить эффективность и качество распознавания речи. Важно отметить еще и то, что на текущий день существует малое количество работ, специализирующихся на распознавании зашумленной русской речи.

Целью настоящей работы является разработка интеграционного метода распознавания русской речи при наличии шума.

II. Технологии распознавания речи, основанные на нейронных сетях

Существует большое количество инструментов и технологий распознавания речи, основанные на нейронных сетях. К ведущим решениям с открытым исходным кодом можно отнести Mozilla DeepSpeech и Kaldi. Все методы делятся на две группы: интегральные и гибридные. Гибридные решения состоят из множества отдельных компонентов, ошибка в одном компоненте может привести к проблемам в других и повлиять на общий результат (качество распознавания). Создание гибридных решений сложнее,

Таблица I
Сравнение различных моделей распознавания речи [4]

Модель	Технология	Речевой корпус	WER %
Гибридные СММ/ИНС модели			
CNN	Torch7	WSJ (Nov'92)	6.7
Kaldi-dnn5b-pretrain-dbn-dnn-smbr recipe	Kaldi	WSJ (Nov'92)	3.35
CTC модели			
RNN-CTC + Kaldi + trigram LM	Kaldi	WSJ (Nov'92)	6.7
LSTM-CTC + trigram LM	Eesen	WSJ (Nov'92)	7.9
Шифратор-дешифратор модели			
CNN + RNN + CTC	Baidu	WSJ (Nov'92)	4.42
CNN + ASG	Torch7,Baidu	LibriSpeech	7.2

чем создание решений, основанных на интегральном подходе: каждый компонент системы необходимо подбирать и настраивать под конкретную задачу. Интегральный метод заключается в создании одной нейронной сети, которая не нуждается в других компонентах, таких как акустическая или языковая модели К недостаткам такой модели можно отнести большой размер обучающей выборки.

III. Сравнительный анализ существующих реализаций систем автоматизированного распознавания речи

A. Метрики оценки систем автоматизированного распознавания речи

Корректная оценка результатов работы систем автоматизированного распознавания речи (англ. automatic system recognision, далее ASR системы), и как следствие возможность корректно сравнить разные ASR системы, имеет большое значение как для конечных пользователей, так и для разработчиков таких систем. В данной работе представленные метрики будут использоваться не только для сравнения систем, но и для оценки конечного результата работы предложенного метода. Для ASR систем существует две основные группы метрик оценивания [5]:

- метрики точности распознавания;
- метрики скорости распознавания.

Основным способом оценки точности распознавания являются метрики, основанные на расстоянии Левенштейна [6]. Расстояние Левенштейна — это метрика, определяющая разницу между двумя символьными последовательностями. Она рассчитывается как количество операций удаления, вставки и замены преобразовывающих одну последовательность символов в другую. Наиболее распространенными метриками, основанными на расстоянии Левенштейна, являются WER - количество ошибочных слов в предложении и SER количество ошибочных предложений.

Важным параметром любой системы является скорость ее работы. Для ASR метрикой, на основе которой вычисляется скорость работы, является метрика SF(RT). Она считается как отношение скорости обработки аудиофайла к длительности этого аудиофайла. К примеру, если файл длительность в одну минуту обрабатывается тридцать секунд, то SF = 0.5.

Естественным условием для сравнения разных ASR-систем с помощью этой метрики является запуск тестов на одинаковом оборудовании

B. Сравнительный анализ ASR по метрике WER

Проведём сравнительный анализ различных моделей для распознавания речи, по трём основным группам: гибридные СММ/ИНС модели, CTC-модели, шифратор-дешифратор модели на основе механизма внимания. Возьмём только две лучше модели в каждой группе по показателю WER.

Гибридные СММ/ИНС модели состоят из блока скрытых марковских моделей(СММ) определяющего наиболее вероятную последовательность фонем и блока искусственной нейронной сети(ИНС), которая вычисляет вероятность последовательности полученную от СММ.

CTC (Connectionist Temporal Classification) позволяет моделям рекуррентных нейронных сетей обучаться без начального выравнивания звуковой дорожки и транскрипции.

Шифратор-дешифратор модели используются для задач, где длины входной и выходной последовательностей являются переменными. Шифратор это нейронная сеть, которая выделяет признаки из входного сигнала в промежуточное представление. Дешифратор это рекуррентная нейронная сеть, которая использует промежуточное представление для генерации выходных последовательностей.

Для сравнения моделей были использованы два набора англоязычных аудиозаписей WSJ(Nov'92)[7] и LibriSpeech[8]. Оба этих набора являются стандартами для тестирования англоязычных ASR.

Как можно увидеть из Таблицы I. однозначными лидерами по показателю WER являются технологии Kaldi и Baidu. Далее мы будем использовать их реализации: vosk - реализацию Kaldi для русского языка и DeepSpeech - открытый проект компании Mozilla реализующий технологию Baidu.

Даже на этой небольшой выборке видно насколько обширно количество способов настройки моделей машинного обучения, и на сколько сильно отличаются показатели качества даже в рамках одной технологии. Кроме того, ни одна из этих систем не проводит анализ распознанного текста, так как единицей их работы являются морфемы - т.е. звуки. Предложенный интеграционный метод предпо-

лагает получение результатов от нескольких разных ASR-систем, и проводит коррекцию ошибок основываясь на результатах других ASR, подбирает наиболее вероятные слова там, где распознание не удалось. Выбор из полученных вариантов может произвести оператор-человек или система контекстного анализа.

Другими словами, предлагается сделать ансамбль ASR систем с коррекцией ошибок.

IV. Интеграционный подход распознавания зашумленной речи

A. Описание метода

Основные этапы предложенного подхода: проводится очистка аудиозаписи от шумов, после этого выполняется ее распознавание с помощью нескольких разных систем автоматического распознавания речи. Полученные результаты составят список наиболее вероятных гипотез (N-Best-List)[12], выбор из которых может произвести либо оператор-человек, либо система контекстного анализа.

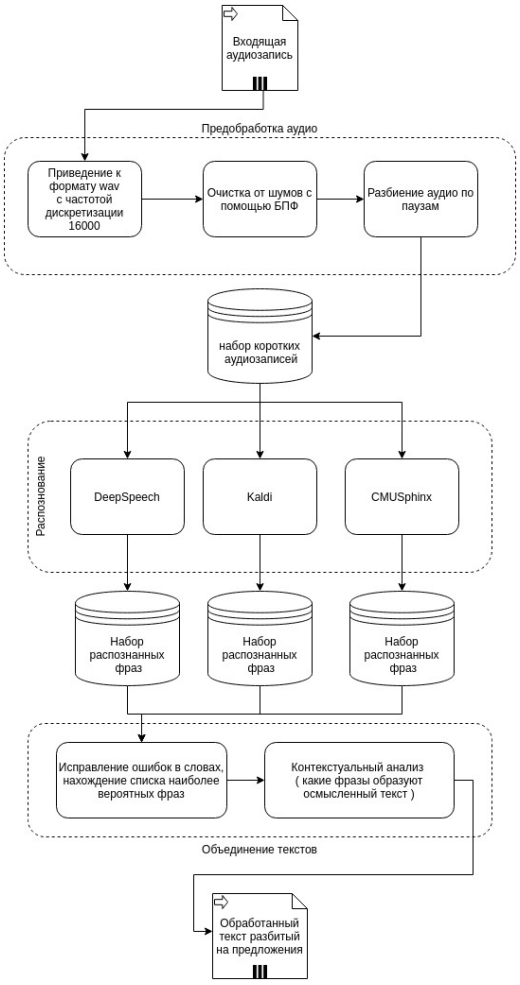

Рис. 1. Структурная схема предлагаемого подхода

На вход программной системы передается аудиозапись, после прохождения нескольких этапов на выходе пользователь системы получает наиболее вероятное предложение.

Первый этап, предобработка аудио. Аудиозапись приводится к заданному формату с конкретной частотой дискретизации. Затем производится очистка от шумов, например с помощью быстрого преобразования Фурье (далее БПФ). Очищенная от шумов аудиозапись разбивается на более мелкие по паузам, тем самым решается несколько проблем: во-первых, мы заранее знаем где были паузы - т.е. законченные мысли и можем это использовать при выдаче конечного результата, во-вторых, мы частично избежим проблемы смешения дикторов.

Второй этап, распознавание аудио. Полученные аудиозаписи помещаются в базу данных и маркируются как относящиеся к одному тексту. Каждая аудиозапись отправляется параллельно во все системы ASR, на выходе которых мы получаем варианты распознанной фразы. После обработки всех аудиозаписей и получения наборов распознанных фраз можно приступать к анализу текстов.

Третий этап, коррекция ошибок. Сначала мы исправляем ошибки в каждой фразе - сравнивая её с вариантами от других ASR, и составляем наиболее полное предложение. Затем в этом предложении проводится обработка последовательностей, разделенных пробелами - мы определяем является ли последовательность словом, если нет, то какие варианты слов из алфавита могут ей соответствовать. Если последовательность невозможно распознать она маркируется спецсимволом MASK.

Четвертый этап, коррекция ошибок на основе контекста. На данном этапе с помощью ручного или автоматического анализа контекста выявляется, составляют ли полученные фразы осмысленный текст. Автоматический анализ контекста предлагается производить с помощью BERT.

B. Алгоритм

Предложенный метод призван уменьшить количество ошибок и как следствие повысить качество распознавания речи. Сделать это предлагается за счет уменьшения пространства всех возможных фраз, путем использования нескольких распознающих систем и получения нескольких возможных вариантов фраз, из которых и будет производится дальнейший выбор.

Полученные фразы должны быть сопоставлены, выявление наиболее вероятных вариантов фраз происходит по следующему алгоритму (описан для трех систем).

- проверяем не состоит ли фраза из одного слова;
- удаляем все пробельные символы и определяем является ли получившийся результат словом с заданным редакционным расстоянием;
- если предыдущий пункт верен, обработку можно считать завершенной.

Все три варианта фразы сортируются по следующим параметрам:

- совпадения количества пробельных символов у нескольких фраз, это свидетельствует о правильном определении границ слов;
- совпадения длины строки;

- по количеству точно распознанных слов (сколько слов из фразы есть в словаре);
- по приоритету ASR, если по какой-то причине мы доверяем одной из ASR больше.

После сортировки принимаем первую фразу за истинную. Выравниваем фразы, по совпадающим словам, заменяя пробелы вокруг совпавших слов на спецсимволы. Таким образом мы получаем границы, правильно распознанных участков.

Промежутки, находящиеся внутри спецсимволов, сравниваем по описанному выше алгоритму, не совпавшие промежутки обозначаем как не распознанные. Если ни в одном промежутке из группы нет хотя бы одного корректного слова, помечаем этот диапазон как не распознанный.

Таким образом, получаем лучшую из возможных комбинацию результатов, в которой не распознанные участки помечены спецсимволом MASK. Если предложение не удалось распознать полностью, то фраза анализируется с помощью BERT - анализатора контекста от компании Google.

C. Особенности очистки от шума

Одной из задач, которую необходимо было решить в рамках данной работы, является задача предобработки звука и удаления шумов. Нам необходимо это сделать не только для уменьшения вероятности ошибки, но и для большей однородности записей.

Есть два основных способа решения этой проблемы: модели на основе рекуррентных нейронных сетей, и различные алгоритмы спектрального анализа. В работе был произведен сравнительный анализ двух инструментов, реализующих эти подходы: RNNoise и ffmpeg.

В рамка проекта ffmpeg разработан фильтр afftdn, предназначенный для очистки аудио от шума. В основе этого фильтра лежит алгоритм БПФ.

RNNoise — это свободный инструмент, основанный на рекуррентной нейронной сети с типом ячеек GRU. Модель RNNoise обученная на различных видах шумов, пытается анализировать аудиозапись и вычленять различные виды шума.

При практическом использовании оказалось, что фильтр afftdn справляется с задачей лучше RNNoise, и работает быстрее, поэтому для очистки шумов был выбран именно он.

D. Реализация подхода

Для реализации предложенного подхода разработана программная система, отвечающая следующим требованиям:

1) возможность встраивания новых процедур для обработки аудиозаписи;
2) возможность параллельного распознавания аудиозаписей с помощью ASR, следовательно необходима возможность одновременного использования одного аудио файла;
3) удобство использования и замены разных ASR использующих разные библиотеки;
4) предоставление кроссплатформенного интерфейса для работы с системой.

Для обеспечения перечисленных требований используется ансамбль докер контейнеров, задачи которым устанавливаются через REST-API сервис, который выступает интерфейсом для внешних пользователей и выполняет функцию брокера задач используя очередь задач.

REST-API сервис реализует архитектуру приложения “Клиент-Сервер“, тем самым обеспечивает кроссплатформенность системы. Использование REST-сервиса предоставляет широкие возможности для клиентских приложений. Клиентское приложение может быть написано под практически любую операционную систему, под практически любую платформу (включая мобильную), и практически на любом языке программирования. Вместо аутентификации пользователя REST-API сервис реализует простой способ защиты данных пользователей. При загрузке файла клиент получает уникальный ключ, полученный на основе переданной аудиозаписи, вычисленный с использованием хеш-функции. Результат распознавания предоставляется в ответ на получение этого ключа. Для предотвращения перехвата ключа, связь между клиентом и сервером реализована через протокол HTTPS.

Для того чтобы повысить отклик программной системы каждый из перечисленных сервисов должен иметь возможность обрабатывать аудиозаписи независимо. Это сложно реализовать при обработке всей аудиозаписи целиком, поэтому решено разбивать аудиозапись на меньшие отрезки ориентируясь на паузы в речи. По умолчанию, разбиение осуществляется на группы, в которых встречается 100 пауз по 2 секунды. Эти два параметра (длина интервала паузы и количество пауз) настраивается в системе для гибкого конфигурирования и получения наилучшего качества распознавания. Чтобы сервисы имели возможность одновременно работать с одним и тем же участком аудиофайла каждый участок загружается в сервис хранения файлов откуда любой другой сервис может его получить. Разбиение позволило не только использовать независимость сервисов, но и снизило время ожидания пользователем первого предложения. Независимость работы всех компонентов позволила использовать все ASR-системы одновременно, тем самым обеспечив параллельность системы на этапе распознавания аудиозаписи.

В качестве конкретных инструментов для реализации были выбраны:

1) язык разработки - Python3.6;
2) очередь задач - Redis;
3) сервис конвертации и очистки от шума - ffmpeg;
4) сервис диаризации - pyAudioAnalysis;
5) сервис хранения файлов - Scality S3;
6) сервисы ASR - Kaldi, DeepSpeech, CMUSphinx.

E. Результаты

Результаты работы программной системы были оценены по показателям WER и SF, описанных в пункте III-A. Тестовые записи, взятые из набора русскоязычных аудио open_stt, обладают следующими характеристиками:

- 15 минут записи - 1400 слов;
- 17Мб;
- 128 Kbit/sec;
- 3% шума.

Замеренные показатели сравнивались со средним значением трех ASR систем, лежащих в основе программной системы. На проверочном наборе данных предложенная система показала ухудшение по показателю SF в среднем на 27%. Другими словами, программная система работает медленнее, что объясняется большим количеством компонентов. По показателю WER программная система показала результаты лучше на 7%, снизив количество ошибок за счет контекстного анализа.

V. Заключение

В работе приведен обзор методов преобразования аудиозаписей в текст. Проведен сравнительный анализ существующих реализаций для рассмотренных методов, на основе которого сделан вывод что интегральные системы пока что немного уступают в точности распознавания гибридным CMM/ИНС моделям.

Представлен интеграционный подход, который комбинирует различные ASR с системами улучшения аудио и обработкой текста.

Приведены детали реализации и проведен анализ результатов по двум метрикам качества, который показывает выигрыш используемого метода над существующими подходами.

Список литературы

[1] Using the Doc2Vec Algorithm to Detect Semantically Similar Jira Issues in the Process of Resolving Customer Requests Kovalev, A., Voinov, N., Nikiforov, I. 2020 Studies in Computational Intelligence

[2] Федеральный закон от 27.07.2006 n 152-фз (ред. от 31.12.2017) "О персональных данных"

[3] Балакшин Павел Валерьевич. Алгоритмические и программные средства распознавания речи на основе скрытых марковских моделей для телефонных служб поддержки клиентов: диссертация кандидата технических наук: 05.13.11 / Балакшин Павел Валерьевич;[Место защиты: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»].- Санкт-Петербург, 2015.- 127 с.

[4] Марковников, Н. М., и И. С. Кипяткова. Аналитический обзор интегральных систем распознавания речи. Труды СПИИРАН, т. 3, вып. 58, June 2018, сс. 77-10, doi:10.15622/sp.58.4.

[5] Карпов Алексей Анатольевич, Кипяткова Ирина Сергеевна. "Методология оценивания работы систем автоматического распознавания речи"Известия высших учебных заведений. Приборостроение, vol. 55, no. 11, 2012, pp. 38-43.

[6] Прытков В.А. "Функция расстояния между строками на основе кусочно-постоянной модели"Доклады Белорусского государственного университета информатики и радиоэлектроники, no. 4 (74), 2013, pp. 22-28.

[7] Paul, Douglas B. and Janet M. Baker. “The design for the wall street journal-based CSR corpus.” ICSLP (1992).

[8] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books,"2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 5206-5210.

[9] Makovkin K.A. [Hybrid models – Hidden Markov Models/Multilayer perceptron and their application in speech recognition systems. Servey]. Rechevye tehnologii – Speech Technology. 2012. vol. 3. pp. 58–83. (In Russ.).

[10] Markovnikov N.M., Kipyatkova I., Karpov A., Filchenkov A. Deep neural networks in Russian speech recognition. Proceedings of 2017 Artificial Intelligence and Natural Language Conference. 2017. pp. 54–67

[11] Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady. 1996. vol. 10. pp. 707–710.

[12] Yen-Lu Chow and Richard Schwartz. 1989. The N-Best algorithm: an efficient procedure for finding top N sentence hypotheses. In Proceedings of the workshop on Speech and Natural Language (HLT ’89). Association for Computational Linguistics, USA, 199–202. DOI:https://doi.org/10.3115/1075434.1075467

[13] Ronzhin A.L., Karpov A.A., Li I.V. Rechevoj i mnogomodal’nyj interfejsy [Speech and multimodal interfaces]. M.: Nauka. 2006. 173 p. (In Russ.).

[14] Kipyatkova I., Karpov A. DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi. International Conference on Speech and Computer. 2016. pp. 246–253.

[15] LeCun Y., Bengio Y. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. 1995. vol. 3361. no. 10. pp. 1995.

[16] Романенко А.Н., Матвеев Ю.Н., and Минкер В.. "Перенос знаний в задаче автоматического распознавания русской речи в телефонных переговорах"Научно-технический вестник информационных технологий, механики и оптики, vol. 18, no. 2, 2018, pp. 236-242.

[17] Povey D. et al. The Kaldi speech recognition toolkit». IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society. 2011. 4 p.

[18] Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx) Këpuska V, Bohouta G

[19] Zaity B., Wannous H., Shaheen Z., Chernoruckiy I., Drobintsev P., Pak V. "A hybrid convolutional and recurrent network approach for conversational AI in spoken language understanding."(2019).

Integration approach for automatic speech recognition of noisy Russian language

Daniil Gomonyuk, Igor Nikiforov, Dmitry Drobintsev

Higher School of Software Engineering

Peter the Great St. Petersburg Polytechnic University

St. Petersburg, Russia

***Abstract*—The research considers methods for the automated conversion of audio recordings into a text data format, in other words, speech recognition. Particular emphasis is placed on the recognition of noisy Russian-language speech.**

The paper provides an overview of existing speech recognition methods which include end-to-end and modular methods. There is a review and comparative analysis of existing implementations of the methods and their metrics. Based on a comparative analysis, it is concluded that Mozilla DeepSpeech technology is the most powerful speech recognition tool.

A distinctive feature of the work is the use of the combined recognition method, which allows to improve the recognition quality of noisy recordings. The end-to-end and modular methods are combined in single approach. The proposed approach is implemented in a software package for recognizing noisy Russian-language speech using Mozilla DeepSpeech technology. The results showing the effectiveness of the proposed end-to-end method are demonstrated.

The developed software package can be used in companies engaged in technical support and call centers to improve the efficiency of processing customer requests.