УДК 004.032.26

Ахметшин А.Ф.

Студент

Казанский государственный энергетический университет

Россия, Казань

Научный руководитель: Насыров И.К.

Доктор технических наук

Казанский государственный энергетический университет

Россия, Казань

ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ ДЛЯ АНАЛИЗА РЕЧЕВОЙ ИНФОРМАЦИИ

Аннотация: Данная статья посвящена использованию нейронных сетей в распознавании речи. Представлена общая схема анализа и обработки речи, систематизирован и приведен перечень требований, предъявляемых к нейросетевым системам анализа и обработки речи. Описаны методики оценки результатов работы систем распознавания речи, обозначены недостатки существующих нейросетевых технологий.

Ключевые слова: распознавание речи, нейронные сети, нейросетевые технологии

Akhmetshin A.F.

Student

Kazan State Power Engineering University

Russia, Kazan

Scientific adviser
Nasyrov I.K.
doctor of engineering
Kazan State Power Engineering University
Russia, Kazan

APPLICATION OF NEURAL NETWORK TECHNOLOGIES FOR SPEECH INFORMATION ANALYSIS

Annotation: This article is devoted to the use of neural networks in speech recognition. The general scheme of speech analysis and processing is presented, the list of requirements for neural network systems of speech analysis and processing is systematized and given. Methods for evaluating the performance of speech recognition systems are described, and disadvantages of existing neural network technologies are identified.
Keywords: speech recognition, neural network, neuronet technology

Введение

Всё чаще пользователь общается не с человеком-оператором, а с вычислительной техникой. Используя технологии распознавания речи, проще сделать тот или иной запрос, а также получить необходимую информацию.

Существуют десятки различных систем анализа и обработки речи, в том числе с помощью нейросетевых технологий. Основную задачу – распознавание речи – они решают с разной степенью эффективности. Эффективность определяется степенью тождества полученного результата и исходного материала. Для повышения эффективности необходимо, чтобы система распознавания речи удовлетворяла определенным требованиям. В статье систематизирован и проанализирован перечень

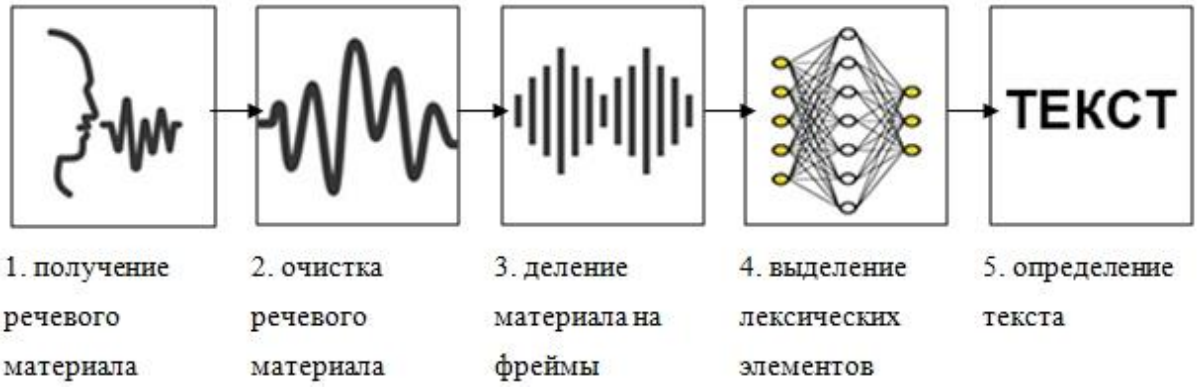
требований, предъявляемых к нейросетевым системам анализа и обработки речи.

Целью настоящей статьи является систематизация основных требований к нейросетевым системам анализа и обработки речи. Для достижения цели необходимо решить следующие задачи: определить общую схему работы системы распознавания речи, выявить недостатки в нейросетевых технологиях обработки речевого материала и возможные пути их решения.

Нейросетевые системы распознавания речи и основные требования к ним

Все нейросетевые системы анализа и обработки речи работают по схожей схеме (схема 1). На первом этапе перед распознаванием речевого сигнала происходит его обработка. В ходе этого процесса удаляются шумы и посторонние сигналы, частотный спектр которых находится вне спектра человеческой речи. Затем сигнал делится на фреймы, из которых извлекаются наиболее важные признаки. Нейронные сети выделяют такие лексические элементы речи как фонемы и аллофоны. Это первый уровень распознавания. На следующих уровнях выделяются слоги и морфемы, а дальше – слова, предложения и сообщения.

Схема 1.



Несмотря на серьезный прогресс в области распознавания речи, стопроцентно эффективной системы еще нет.

На сегодня актуальны проблемы дикторонезависимости и помехоустойчивости. Современные системы распознавания речи, позиционируемые как дикторонезависимые, распознают изолированные слова. Как правило, их словарь ограничен и точность распознавания в таких системах около 95%. Для анализа и обработки слитной речи требуется более обширный словарь и тонкая настройка на конкретного диктора. В таких системах, единицей распознавания на акустико-фонетическом уровне, обычно является аллофоны, дифоны, фонемы и другие фонемоподобные элементы языка. Эффективность таких нейросетей относительно низкая, особенно при анализе и обработке речевого сигнала нескольких дикторов, каждый из которых требует своих настроек нейросети.

Каждый человек разговаривает по-своему. При этом одно и то же слово он может произносить по-разному. Кроме того, у одного и того же человека может меняться скорость речи. Особую сложность для речевых систем представляют региональные диалекты и акцент.

Также проблему для компьютера представляют омонимы, слова с одинаковым звучанием, но с разным значением. Чтобы выбрать подходящий смысл, программа распознавания речи должна анализировать контекст.

Другая проблема – помехоустойчивость – может решаться по нескольким направлениям. С одной стороны, необходимо устранять помехи, шумы и искажения, влияющие на правильное распознавание речи. С другой стороны, существует возможность выделения требуемого речевого сигнала из всего многообразия акустической среды.

Как правило, работа ведется одновременно в нескольких направлениях. Системы, с высокой степенью отвечающие требованиям дикторонезависимости и помехоустойчивости, презентовали такие крупные компании как Google, Alibaba. Разработчики Google запустили

проект VoiceFilter для идентификации личности человека по голосу в толпе. Для его реализации потребовалась одновременная работа сразу двух нейросетей — для распознавания говорящего и для сравнения звуковых спектрограмм, имеющихся в базе данных. В проекте Alibaba искусственный интеллект обрабатывает речь в режиме онлайн, используя облачный сервис. Алгоритм нейросети взаимодействует с системой направленно-удалённых микрофонов, что позволяет отрезать лишние шумы.

Одним из важных показателей эффективности системы является скорость обработки речи. Скорость обработки вычисляется по формуле 1.

Формула 1.

$$SF = \frac{RT}{D}$$

где SF (Speed Factor) – скорость обработки речи, RT (Real Time) – время, потраченное на обработку речи, D (duration) – продолжительность обрабатываемого аудиофайла.

Чем меньше показатель SF, тем скорость обработки речи выше.

Другим показателем скорости распознавания речи может быть период ожидания обработки отсчета (SPL – Sample Processing Latency). Этот показатель означает максимальное количество аудиоданных, которое алгоритм распознавания должен обработать до выдачи результата о первом отсчете сигнала.

Для оценки системы распознавания речи используется метрика WER (word error rate – частота ошибок в словах). Метод определения показателя WER состоит в выравнивании двух текстовых строк (первая — это результат распознавания, а вторая — запись того, что было сказано в действительности) с помощью алгоритма динамического программирования с вычислением расстояния Левенштейна. Расстояние Левенштейна представляет собой минимальное количество или

взвешенная сумма операций редактирования для преобразования первой строки во вторую с наименьшим числом операций ручной замены (S), удаления (D) и вставки (I) слов. WER рассчитывается по формуле 2:

Формула 2.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

где S – количество замен, D – количество удалений, I – количество вставок, C – количество правильных слов, N – количество слов.

Чем меньше операций требуется для приведения к тождеству результата распознавания и фразы, произнесенной в действительности, тем более качественной является система распознавания речи.

Эффективность наиболее распространенных систем автоматического распознавания речи с открытым исходным кодом можно увидеть в таблице 1.

Таблица 1.

Система	**WER, %** (чем меньше, тем лучше)	**SF** (чем меньше, тем лучше)
HTK	19,8	1.4
CMU Sphinx (pocketsphinx/sphinx4)	21.4/22.7	0.5/1
Kaldi	6.5	0.6
Julius	23.1	1.3
iAtros	16.1	2.1
RWTH ASR	15.5	3

С каждым днем нейронные сети всё лучше и лучше обрабатывают и анализируют речевой материал. Растет количество данных, на которых нейронные сети учатся распознавать речь. Несомненно, что уровень распознавания речи будет повышаться и возможно через несколько лет пользователь перестанет понимать, кто с ним общается – человек или нейронная сеть.

Заключение

В статье систематизированы и проанализированы требования к системам анализа и обработки речи. Процесс распознавания речи заключается в том, чтобы выделить, классифицировать и соответствующим образом отреагировать на человеческую речь из входного звукового материала. Это может быть и выполнение определенного действия на команду человека, и выделение определенного слова-маркера из большого массива речи, и системы для голосового ввода текста. С этими задачами нейронные сети справляются всё успешнее.

Для высокоэффективного использования система распознавания речи должна отвечать следующим требованиям:

– высокая скорость обработки речи;

– помехоустойчивость;

– дикторонезависимость.

Исходя из вышеизложенного предлагается следующее. Для решения задач анализа речевой информации, в первую очередь, необходимо одновременное использование нескольких нейросетей с четким разграниченным функционалом. Одна нейросеть очищает речевой материал от помех, другая – распознает слова, третья – находит соответствия в базах данных. Это позволит не только повысить качество распознавания речи, но и даст возможность точнее регулировать механику работы конкретной нейросети, не влияя на работу других сетей.

Во-вторых, задачу помехоустойчивости можно решить с помощью машинного обучения и использования нескольких больших баз данных с речевым материалом разной степени чистоты. Также предлагается применение баз данных шумов для более эффективной очистки речевого материала.

Наиболее сложным для реализации является процесс распознавания речи без настройки на диктора. Необходимо, чтобы система распознавала любое включенное в словарь слово, кем бы оно ни было произнесено. Для

решения этой задачи разработчикам приходится опрашивать большое число (несколько сотен или тысяч) носителей языка, выделять некие общие элементы речи, усреднять их – и все этого для того, чтобы обеспечить распознавание нескольких десятков слов. Чаще всего словарь без настройки на голос пользователя требует раздельного произнесения слов. В этой связи одним из вариантов решения задачи видится использование техники поиска ключевых слов и анализ контекста. Нейросеть должна не только переводить речь в текст, но и анализировать его. Определяя тематику и контекст речи, можно исключить часть словаря и повысить качество распознавания исходного материала.

Удовлетворение этих требований позволит создать систему анализа и обработки речи с высоким качеством распознавания речевого материала.

Использованные источники:

1) Карпов А.А. Методология оценивания работы систем автоматического распознавания речи / А.А. Карпов, И.С. Кипяткова // Известия высших учебных заведений. Приборостроение. – 2012. – Т. 55. – №. 11. – С. 38-43.

2) Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. – Доклады Академий Наук СССР, 1965. – С.12-25.

3) Гусев М.Н. Система распознавания речи: основные модели и алгоритмы / М.Н. Гусев, В.М. Дегтярев. – СПб.: Знак, 2013. – 128 с.

4) Беленко М.В. Сравнительный анализ систем распознавания речи с открытым кодом / М.В. Беленко, П.В. Балакшин // Международный научно-исследовательский журнал. — 2017. — № 04 (58) Часть 4. — С.13—18.

5) Ле, Нгуен Виен. Распознавание речи на основе искусственных нейронных сетей / Нгуен Виен Ле, Д. П. Панченко. — Текст: непосредственный // Технические науки в России и за рубежом: материалы

I Междунар. науч. конф. (г. Москва, май 2011 г.). — М.: Ваш полиграфический партнер, 2011. — С. 8-11.