РАЗРАБОТКА СИСТЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ БЕЛОРУССКОЙ РЕЧИ

С. А. Омелянюк

Белорусский государственный университет, г. Минск;
omse1998@yandex.by; науч. рук. – А. В. Колесов

В работе рассматриваются методы построения систем для автоматического распознавания речи. Цель работы – разработка такой системы для речи на белорусском языке и оценка её эффективности. В результате была получена модель, показывающая многообещающие результаты на различных доменах. Было проведено сравнение качества разработанной системы с имеющимися в открытом доступе системами распознавания речи для белорусского языка. В процессе работы также был собран и обработан корпус данных для данной задачи, построена статистическая языковая модель для белорусского языка.

Ключевые слова: автоматическое распознавание речи; белорусский язык; машинное обучение; нейронные сети.

РАСПОЗНАВАНИЕ РЕЧИ

Задача распознавания речи состоит в том, чтобы по последовательности звуков из аудиозаписи определить транскрипцию речи, произнесённой в ней. Типичная система распознавания речи включает в себя акустическую и языковую модели. Акустическая модель моделирует отношение между аудио сигналом и единицами распознаваемого алфавита. Языковая модель оценивает вероятность появления в языке некоторого предложения.

Весь входной звук разбивается на маленькие части, называемые фреймами. На каждом из фреймов акустическая модель предсказывает распределение символов алфавита. После этого, на основе информации от акустической и языковой моделей происходит процесс декодирования, в результате которого система вырабатывает гипотезы распознавания – наиболее вероятные транскрипции.

Акустическая модель

В качестве акустической модели в данной работе использовалась глубокая нейронная сеть на основе архитектуры TDNN [1]. Обучалась данная сеть при помощи функции потерь CTC [2]. Чтобы улучшить качество модели и уменьшить вероятность её переобучения под корпус малого размера, была использована техника трансферного обучения. А именно, для обучения модели распознаванию белорусской речи была

взята готовая модель для распознавания речи на русском языке и переобучена путём замены классифицирующего слоя сети.

Языковая модель

Наиболее широкое распространение получили статистические языковые модели. Они предсказывают вероятность появления следующего слова на основе предыдущего контекста ограниченного размера. Строятся такие модели на основании текстового корпуса. Модель является n-граммной, если при её построении количество слов в контексте равно $(n-1)$.

Для построения четырёхграммной языковой модели в данной работе был собран корпус на основе текстов из белорусских книг в открытом доступе, всей белорусской секции Википедии и новостей с сайта агентства БелТА.

Метрика качества

Наиболее часто используемая метрика качества в распознавании речи – WER (англ. Word Error Rate). Вычисляется она следующим образом. Для начала, подсчитывается пословное расстояние Левенштейна между исходным предложением и гипотезой распознавания. Затем значение расстояния делится на количество слов в исходном предложении. Приближённо, данная метрика оценивает долю неправильно распознанных системой слов.

ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Корпус данных

Тренировочный корпус был получен на основании пяти аудиокниг, взятых с электронного ресурса «Беларуская Палічка» [3]. Аудиокниги были подобраны так, чтобы дикторы в них были различными, иначе бы происходило переобучение акустической модели под звуковые характеристики одного диктора. Для сбора корпуса было разработано приложение, позволяющее в полуавтоматическом режиме сегментировать данные из аудиокниг. Всего было собрано около 10 часов данных.

Для оценки качества модели было собрано несколько наборов данных из разных акустических доменов: аудиокниги, те же, что в тренировочном корпусе; речь из выпуска новостей; речь, зачитанная на мобильный телефон; диалоги из театральной постановки.

Все числительные в корпусе были заменены текстом, удалена пунктуация, все буквы приведены к строчному виду.

Анализ результатов

Основные результаты представлены в Таблице 1. В ней производится сравнение полученной модели по метрике качества с моделью, разработанной в Лаборатории распознавания и синтеза речи ОИПИ НАН РБ [4]. Несмотря на значительную разницу в качестве, доля неправильно распознанных слов остаётся достаточно большой. Это можно объяснить тем, что акустический домен данных из тренировочного корпуса не совпадает с несколькими доменами тестового корпуса.

Таблица 1

Качество распознавания речи по метрике WER, %

Модель	Аудиокниги	Новости	Мобильные данные	Театральная постановка
Лаборатория распознавания и синтеза речи	91.81	98.7	88.8	97.47
Данная статья	25.65	66	46.19	68.94

На Рисунке 1 представлена тепловая карта распределения вероятностей в логарифмической шкале каждого из символов распознаваемого алфавита в каждый момент времени. По ней можно сделать вывод, что модель уже достаточно хорошо распознаёт часть символов. Буквы в ячейках карты означают расположение оптимального выравнивания исходного предложения согласно алгоритму динамического программирования.

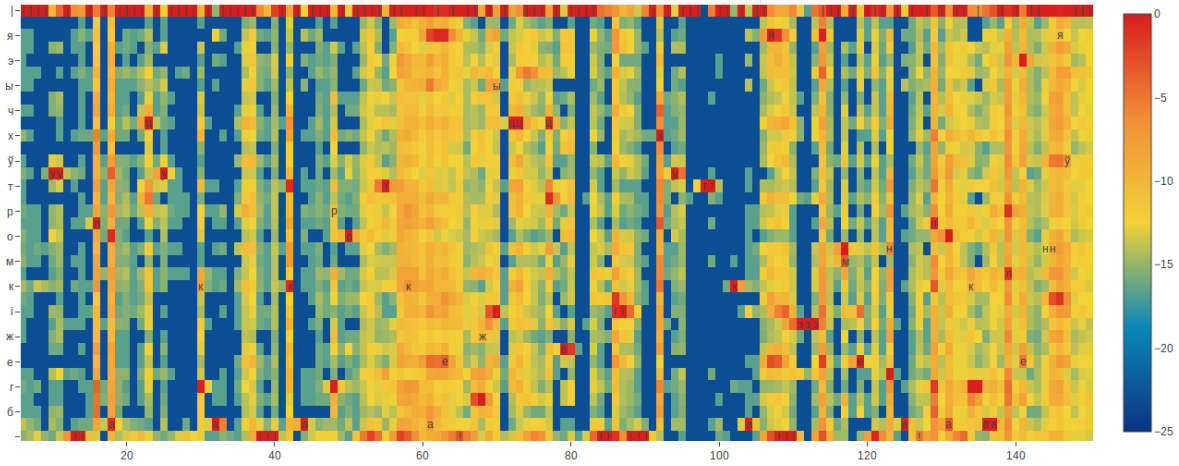

Рис. 1. Тепловая карта выходов акустической модели

В Таблице 2 приведены примеры исходного текста и результата работы системы распознавания речи.

Таблица 2

Примеры результата работы системы распознавания

Исходное предложение	Гипотеза распознавания
шэрсць бліскучая сам вясёлы усе органы і залозы функцыяніруюць нармальна	сэр бліскучая сам вясёлых узе органы залозы панцыяні нарва
ці няма ў вас якіх небудзь доказаў	ці нямала якіх небудзь даказаў
магу захапляцца і рамантыкай але свае рамантычныя захапленні заўсёды правяраю статыстычнымі данымі	магу захапляцца рамантыкі але свое рамантычнае захапленне заўсёды правяла з тага стэна вядомымі
у пацука кароткае жыццё і хуткая змена пакаленняў	у пацука таго т віццё і хуткая з не дарога

ЗАКЛЮЧЕНИЕ

Таким образом, в результате проделанной работы была разработана система распознавания белорусской речи, показывающая удовлетворительные результаты по метрикам качества. Эксперименты показывают, что модель может хорошо справляться с распознаванием речи на акустических доменах, схожих с присутствующими в тренировочном корпусе, а также показывать неплохие результаты на более отдалённых доменах.

Библиографические ссылки

1. *Peddinti V., Povey D., Khudanpur S.* A time delay neural network architecture for efficient modeling of long temporal contexts //Sixteenth Annual Conference of the International Speech Communication Association, Dresden, 2015. P. 3214–3218.
2. *Hannun A.* Sequence modeling with ctc // Distill. 2017. Vol. 2. №. 11. P. e8.
3. Беларуская палічка : [сайт]. URL: http://www.knihi.com.
4. Thematic Speech Recognizer [Электронный ресурс] // SSRLab UIIP NAS Belarus. URL: https://corpus.by/ThematicSpeechRecognizer (дата обращения: 26.05.2019).