Қазақстан Республикасы Тәуелсіздігінің 30 жылдығына арналған «Сейфуллин оқулары – 17: «Қазіргі аграрлық ғылым: цифрлық трансформация» атты халықаралық ғылыми – тәжірибелік конференцияға материалдар = Материалы международной научно – теоретической конференции «Сейфуллинские чтения – 17: «Современная аграрная наука: цифровая трансформация», посвященной 30 – летию Независимости Республики Казахстан.- 2021.- Т.1, Ч.4 - С.176-179

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ КАЗАХСКОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Серік К.Н.

Создание человеко-машинных интерфейсов на естественном языке и, в частности, систем автоматического распознавания речи в последнее время стало одним из основных направлений и задач в области искусственного интеллекта. Речевые технологии обеспечивают более естественное взаимодействие пользователя с вычислительными и телекоммуникационными комплексами по сравнению со стандартным графическим интерфейсом.

С развитием персональных компьютеров и широкого спектра общественно-информационных и развлекательных услуг, речевые, а затем и мультимодальные интерфейсы теперь больше ориентированы на применение в социальных интеллектуальных сервисах, что накладывает свои оттенки на системы обработки речи. В частности, расширяется словарный запас лексических единиц, увеличивается вариативность речи, а обработка должна осуществляться в режиме реального времени для поддержания естественного диалога с пользователем. Разработка компактного способа представления словаря особенно актуальна для агглютинативных языков с относительно богатой морфологией. Для учета вариабельности и усвоения моделей фонем и слов требуются огромные текстовые и речевые материалы, подготовка которых требует кропотливой экспертной работы.

В данной статье будут проанализированы три вида нарушений речи, наиболее характерных для спонтанной речи: 1) озвученная пауза, 2) повторение слов, 3) модификация предложения с самого начала. В качестве материала использовались речевой корпус Spoken Dutch Corpus (CGN) и Коммутатор-1. Количество озвученных пауз составляло 3 % от всех лексических единиц в этих корпусах. Чаще всего это были междоусобицы, которые располагались во всех частях предложений. Относительное число повторений составило около 1%. А двадцать наиболее частых повторов - это короткие слова, состоящие из одного слога.

Для фильтрации нежелательных речевых сбоев в мультимедийных записях лекций использовался аудиовизуальный детектор озвученных пауз. Записанный мультимедийный корпус лекций продолжительностью около 7 часов содержал изображение экрана планшетного компьютера, на котором лектор делал рукописные заметки, выводимые аудитории на мультимедийный проектор, а также звуковой поток с речью лектора и фоновым шумом. Анализ

тела показал, что подавляющее большинство колебаний происходит, когда лектор не использует планшет (или блокнот, табличку с данными), поэтому для фильтрации пауз использовался двухступенчатый алгоритм. Сначала определялись моменты времени, когда изображение на экране монитора не менялось, а затем только в эти периоды времени осуществлялся поиск заполненных пауз в аудио потоке. При анализе рассматривались озвученные паузы продолжительностью более 120 мс, выраженные как по отдельности (т.е. содержащие сегменты с тишиной до и после колебаний), так и в пределах одного слова. Использование предварительной сегментации звуковых сегментов и видеоанализа с планшета позволило повысить точность распознавания колебаний до 85%.

В условиях бурного развития речевых технологий связано развитие искусственных нейронных сетей и в настоящее время все большую популярность приобретают исследования по использованию DNN для распознавания казахской речи. В то же время, в настоящее время нет эффективных систем автоматического распознавания казахской речи и развитие АСР является актуальным.

В данной статье рассматривается метод создания системы автоматического распознавания казахской речи с использованием DNN с помощью инструментов Калди. В данной работе был расширен существующий речевой корпус, собраны речевой и текстовый корпуса для казахского языка, а также созданы акустические и языковые модели на основе нейронной сети (NN), что позволило повысить точность распознавания казахской речи.

Для препроцессирования речи использовались следующие алгоритмы: Мел-кепстральные коэффициенты (MFCC) и коэффициенты перцептивного линейного предсказания (PLP). Для акустического моделирования используется скрытая марковская модель (HMM), модель гауссовской смеси распределения (GMM), подпространственная гауссовская модель смеси (SGMM) и глубокие нейронные сети (DNN). Языковое моделирование выполняется с использованием конечных преобразователей (FST) с поддержкой линейной алгебры - библиотек BLAS и LAPACK.

В настоящее время DNN часто используется в речевых исследованиях для распознавания речи, и результаты исследований показывают хорошие результаты. Например, исследования представляют систему распознавания спонтанной чешской, словацкой и русской речи для обработки допросов свидетелей Холокоста. В данной работе были созданы базовые транскрипции автоматически по определённому набору правил, а для многих слов было сгенерировано несколько вариантов транскрипции с учётом фонетических феноменов непрерывной речи (например, ассимиляция согласных на границе слова). Затем создавались транскрипции, описывающие варианты произношения, а также для русского языка и акцента, так как интервью брали не только у жителей Казахстана, России, но и у русских, проживающих на Украине, в Израиле, США. Кроме того, моделировались неязыковые явления. Размер корпуса, использованного для создания акустических моделей для

русского языка, составлял 100 часов, использовался DNN. Языковая модель представляла собой биграмматическую модель с использованием метода возврата (схема отступления Катца). При размере словаря 79 тыс. транскрипций процент неправильно распознанных слов составил 38,57%.

Другой класс заявок на распознавание речи - стенография. Чаще всего с такой задачей выполняется некоторый монолог, записываемый в достаточно хороших акустических условиях с помощью микрофона гарнитуры. Поэтому в отличие от систем массового обслуживания, где речь поступает по телефонным каналам и/или записывается на улице, автоматические системы транскрибирования принимают речевой сигнал с гораздо лучшим качеством записи. Поскольку существуют более мягкие требования к скорости распознавания, система может обрабатывать речевой сигнал за несколько проходов, используя методы адаптации к голосу говорящего и применяемую проблему.

Ученые провели исследование по распознаванию непрерывной русской речи с использованием DNN доверия). Для распознавания речи был использован метод с использованием конечных государственных машинных преобразователей. Показано, что предложенный метод позволяет повысить точность распознавания речи по сравнению со скрытыми марковскими моделями.

За последние десять лет в мире было создано несколько речевых корпусов, содержащих до тысячи дикторов, записанных в различных условиях окружающей среды. Запись акустических данных для создания акустического корпуса языка осуществлялась в Институте информационных и вычислительных технологий Научного комитета Министерства образования и науки Республики Казахстан в г. Алматы. Для этого использовалась звукозаписывающая профессиональная студия Vocalbooth.com (рис. 1). Кабина для записи звуковых данных состоит из двух слоев шумоизоляции, с одной и той же герметичной дверью. Дизайн интерьера состоит из пирамидообразного звукопоглощающего акустического материала красного цвета, а кабина оборудована системой бесшумного воздухообмена. Студия предназначена для записи высококачественных звуковых данных.

Рисунок 1 - Звукоизолирующая профессиональная студия звукозаписи компании Vocalbooth.com

В качестве выступающих люди отбирались без проблем с произношением речи. Для целей исследования и дальнейшего использования данных, выступающие опрашивались по ранее созданному шаблону (рис. 1). Для записи было использовано 200 дикторов разного возраста (от 18 до 50 лет) и пола. В среднем на озвучивание и запись одного диктора уходило 40-50 минут. Для каждого оратора был подготовлен текст, состоящий из 100 предложений. Предложения записывались в отдельные файлы. Каждое предложение состоит в среднем из 6-8 слов. Предложения выбираются с использованием наиболее богатой фонемы слов. Текстовые данные собирались с новостных сайтов на казахском языке, другие материалы использовались в электронном виде. Всего 76 часов звуковых данных были записаны. Во время записи создавались транскрипции - описание каждого звукового файла в текстовом файле. Созданный корпус позволяет, во-первых, работать с большими объемами баз данных, проверять предлагаемые характеристики системы и, во-вторых, изучать влияние расширения базы данных на скорость распознавания.

В работе сравниваются языковые модели, построенные с использованием фидфорвд-нейронной сети и рекуррентной нейронной сети. Использованы три различные реализации языковой модели на нейронных сетях: 1) программные средства LIMSI для создания нейронной сети, в которой выходной слой ограничен наиболее часто встречающимися словами; 2) нейронная сеть с кластеризацией (используется весь словарь); 3) рекуррентная нейронная сеть с кластеризацией. Экспериментальные результаты показывают, что языковые модели, построенные с использованием нейросети, работают хуже, чем рекуррентные нейронные сети. По данным тестирования, рецидивирующая сеть показала улучшение на 0,4% по сравнению с использованием нейросети.

Языковая модель - позволяет определить наиболее вероятные последовательности слов. Сложность построения языковой модели во многом зависит от конкретного языка. Так, для английского языка достаточно использовать статистические модели (так называемые N-граммы). Для агглютинативных языков с относительно богатой морфологией статистические модели не подходят и используются гибридные модели.

Языковая модель p(w) дает предварительную вероятность последовательности слов w. В основном она показывает, насколько вероятно произнести последовательность слов, основываясь на грамматических правилах языка. Поскольку эта модель зависит только от текста и не зависит от акустических данных, то в качестве источника входных данных можно использовать большое количество текста, имеющегося в книгах, журналах, статьях и т.п. Кроме того, мы хотим, чтобы языковая модель фиксировала специфическую для данной темы информацию для специальных систем ASR. Для фиксации определенных характеристик, связанных с человеческой речью, например, некоторых грамматических ошибок, часто встречающихся в речи, повторов, колебаний и т.д., транскрипции произнесенного текста также

являются полезным источником входных данных. Поскольку общее количество возможных последовательностей слов не ограничено, для получения достоверных неразборчивых оценок необходимо сделать упрощающие предположения. Стандартным способом вычисления вероятностей языковой модели является накопление количества соседних слов. Он предполагает, что вероятность текущего слова w_n зависит только от предыдущих m-1 слов w_n-1 ... n-m+1.

В данной статье мы рассмотрели систему автоматического распознавания казахской речи, которая работает на базе DNN. По результатам исследования видно, что для автоматического распознавания речи лучше использовать DNN, чем классические алгоритмы. В работе проанализированы существующие модели и методы, рассмотрен алгоритм сжатия речи с использованием алгоритма MFCC. В связи с этим было заявлено, что методы MFCC и DNN дают наилучшие результаты.

Список литературы

1. Schuster, M. Speech Récognition for Mobile Devices at Google / M. Schuster // LNCS. -2010. - Vol. 6230. - P. 8-10.

2. Козлов, В. В. Определение параметров гармонических сигналов в условиях действия шумов и помех на основе метода разложения сигнала на собственные числа / В. В. Козлов // Современные проблемы науки и образования. - 2013. - № 6. - URL: http ://www.science-education.ru/113-10860.

3. Козлов, В. В. Исследование погрешности определения параметров гармонического сигнала на основе метода разложения на собственные числа / В. В. Козлов, Б. Н. Маньжов, Е. А. Ломтев // Измерения. Мониторинг. Управление. Контроль. - 2012. - № 1. -С. 50-55.