**УДК 81.33**
**https://doi.org/10.25076/vpl.36.01**
**И.И. Валуйцева**
**И.Е. Филатов**
**Московский государственный областной университет**

# ПОДХОДЫ К ЯЗЫКОВОМУ И АКУСТИЧЕСКОМУ МОДЕЛИРОВАНИЮ В РАСПОЗНАВАНИИ РЕЧИ

*Исследование посвящено вопросу эффективности традиционных и современных подходов к автоматическому распознаванию речи (ASR). В статье анализируется общая структура алгоритмов машинного распознавания речи, в частности, языкового и акустического моделирования, а также лексических данных; проиллюстрировано историческое развитие автоматического распознавания речи и представлены наиболее современные подходы. Проведен эксперимент, в котором с помощью определённого набора тестовых примеров производится сравнение нескольких приложений для распознавания речи. В выборке присутствует четыре разные системы ASR, основанные на разных алгоритмах акустического и языкового моделирования: во всех компонентах только в двух системах используется один и тот же подход, в двух других приложениях акустические и языковые модели основаны на разных алгоритмах – таким образом, структуры всех элементов выборки принципиально не похожи. Набор данных анализируется каждой системой с помощью программ на языке Python, выходных данные нормализуются и сравниваются по стандарту WER с заранее транскрибированными эталонными данными. Проведен анализ результатов тестирования, сделаны выводы о зависимости эффективности системы автоматического распознавания речи от оптимизации её элементов и обучения при помощи необходимого набора данных, в то время как нейросетевой и статистический подходы являются одинаково результативными в задачах языкового и акустического моделирования.*

*Ключевые слова: распознавание речи, акустическое моделирование, языковое моделирование, скрытые марковские модели, глубокое обучение, нейронные сети.*

**UDC 81.33**
**https://doi.org/10.25076/vpl.36.01**
**I.I. Valuitseva**
**I.Y.Philatov**
**Moscow Region State University**

# METHODS OF LANGUAGE AND ACOUSTIC MODELLING IN SPEECH RECOGNITION

*The research deals with the problem of efficiency of traditional and modern methods of automatic speech recognition (ASR). In the article the analysis of the common machine speech recognition algorithm structure is conducted, particularly language and acoustic models, as well as vocabulary data; the historical development of automatic speech recognition is illustrated, and its most innovative approaches are presented. The authors have conducted an experiment, in which several ASR application programming interfaces are compared with each other with a certain set of test cases. There are four different ASR-systems based on different algorithms of acoustic and language modelling: only two of them apply the same approach for all the elements in their structure, while in the other two applications acoustic and language models are based on different algorithms – thus, the structures of all the elements of the collection are not similar. The data set is analyzed by each system with Python programs; the output data is standardized and compared to the pre-transcribed reference data via WER. The results of the research have been analysed and are leading to a conclusion that the efficiency of ASR-system depends on its elements optimization and learning with an applicable data set, while neural net and statistical methods are both equally useful in acoustic and language modelling tasks.*

*Key words: speech recognition, acoustic modelling, language modelling, hidden Markov models, deep learning, neural nets.*

## Введение

Автоматическое распознавание речи (ASR) – технология, используемая в огромном количестве современных приложений. Одна из целей автоматического распознавания речи – создать

подобие естественного, живого общения между машинами и людьми. На данном этапе развития ASR помогает огромному количеству приложений взаимодействовать с пользователем. Современные системы распознавания и синтеза речи достаточно развиты – используя новейшие технологические разработки, они показывают высочайшие результаты точности работы.

ASR – это полезный инструмент, применяемый в ряде различных разработок. Сегодня на рынке представлено большое разнообразие таких систем – разработкой ASR заинтересованы ведущие технологические компании, такие как Google, Facebook, Microsoft, Weibo, Apple, IBM, Amazon, Яндекс и другие. Действительно, их сервисы, использующие технологии автоматического распознавания речи, известны своей точностью, а сами компании считаются флагманами индустрии.

Несмотря на значительные успехи в области разработки ASR-систем, автоматическое распознавание речи работает не идеально. Рыночная конкуренция заставляет компании развивать ASR, внедряя в них новейшие инструменты, из которых самым популярным и эффективным являются нейронные сети (НС). Разные виды нейронных сетей становятся частью ASR-систем, улучшая точность их работы. Однако эта тенденция не повсеместна – компании ограничивают работу НС внутри сервиса распознавания речи, оставляя часть алгоритмов традиционным методам скрытых марковских моделей, либо отказываются от полного внедрения нейросетей вовсе.

Сегодня на рынке представлено большое разнообразие систем ASR, в основе которых лежат разные алгоритмы работы: от статистических моделей до нейронных сетей. Некоторые их этих систем показывают впечатляющие результаты в точности – менее 10% ошибок при распознавании живой речи, и в каждой из них используется свой подход к обработке языка с помощью акустического и языкового моделирования.

**История инструментов распознавания речи**

Технологии, которые лежат в основе современных систем ASR, развиваются уже на протяжении полувека. Развитие автоматического распознавания речи пережило несколько периодов развития, которые связанны в первую очередь с подходами к обработке естественной речи. Интерес человека к

распознаванию и синтезу речи возник задолго до создания первых ASR-машин, однако первые попытки реализации проектов создания инструментов автоматического распознавания речи произошли лишь в середине двадцатого века. Первые системы распознавания речи могли определять некоторые слова или распознавать отдельные звуки – в основе их работы лежал анализ формант и фонем (Juang & Rabiner, 2004); такие системы не отличались практичностью, но это были первые шаги к созданию полноценных приложений для автоматического распознавания речи.

До 1970-ых годов технологии в области развивались критически медленно – разработчикам не был ясен дальнейший путь развития ASR, и существующие проекты теряли инвесторов. Застой в сфере прекратился в конце 70-ых, когда Министерство обороны Соединенных Штатов открыло проект «Speech Understanding Research», в рамках которого было создано несколько систем распознавания речи – самая успешная из них могла распознать более 1000 слов. Примерно в это же время IBM и AT&T задумываются о разработке коммерческих приложений для автоматической обработки естественной речи – речь шла об автоматических телефонных сетях. Несмотря на достигнутый прогресс, ASR-машины были также далеки от использования среди потребителей.

Поворотной точкой в развитии ASR стал глобальный переход в методологии от интуитивного подхода работы с моделями (парадигма непосредственного распознавания шаблонов) к более точному статистическому методу моделирования – CMM. Хотя концепция работы Скрытых Марковских моделей в автоматическом распознавании речи была уже изучена ранее в нескольких исследовательских группах, подход дорабатывался до середины 1980-ых. Главенствующий в индустрии до сих пор, принцип использования CMM в качестве основы для механизмов автоматического распознавания речи появился именно в те годы; в это же время в сфере возникает понимание того, каким должен быть механизм ASR: машина должна уметь интерпретировать язык, понимать его как систему, научить компьютер «говорить» на человеческом языке невозможно. По большей части развитие ASR-систем с конца 70-х годов двадцатого века заключалось в

совершенствовании компьютеров, обрабатывающих Скрытые Марковские модели (Juang & Rabiner, 2004).

СММ – это двойной вероятностный процесс, моделирующий истинную возможность изменения речевого сигнала (и его признаков) и структуру естественного языка в интегрированной и однородной среде статистического моделирования.

Как известно, реальные речевые сигналы по своему характеру очень различны – люди говорят с разными произношением, акцентами и другими особенностями; в аудиофайлах могут присутствовать шумы. Когда несколько людей произносят одно и то же слово, акустические сигналы их речи значительно отличаются друг от друга, даже если лежащая в основе лингвистическая структура (произношение, синтаксис, грамматика) остается одинаковой. Формализм СММ – вероятностная мера, использующая Марковскую цепь для представления лингвистической структуры и набор вероятностных распределений, отвечающих за изменение акустической реализации звуков.

Вероятностная модель СММ оказалась полезна в задачах декодирования объёмных речевых структур – языковых моделей. Использование конечной грамматики для непрерывного распознавания речи с большим объемом лексических данных представляет собой логичное продолжение концепции использования Марковских цепей, которую СММ применяет для работы с акустическим моделированием. Такой подход эффективно справляется и с языковым моделированием.

Таким образом, Скрытые Марковские модели стали универсальным инструментом, способным обрабатывать речь любого говорящего с огромным набором лексических данных, а их внедрение в механизмы автоматического распознавания речи обозначило важнейший переход ASR от простого распознавания обработанных шаблонов к статистическому методу распознавания естественной речи, используемый и по сей день.

Первые коммерческие продукты ASR были основаны именно на СММ: Dragon Dictate, вышедший в 1990, был самым известным из них. Со словарём объёмом 80000 слов этот инструмент умел обрабатывать 30-40 слов в минуту, что почти в четыре раза медленнее живой речи. Кроме того, Dragon Dictate умел

распознавать только речь, сказанную определённым образом. Но и этого было достаточно для того, чтобы успех сервиса среди покупателей вызвал интерес к разработке похожих систем у других компаний.

**Современные модели распознавания речи**

Современное автоматическое распознавание речи представляет собой процесс поиска лучшей последовательности слов по моделям; модели всего две – акустическая и языковая. Третья сторона системы ASR – лексические данные, под которыми обычно понимается транскрибированный языковой корпус с выделенными фонемами; в прикладных задачах нередко используются специализированные лексические данные (Povey, 2012).

Акустическая модель воспроизводит последовательность векторов признаков при обработке последовательностей звуков. Языковая модель определяет вероятности конкретных последовательностей элементов (слов или морфем). Таким образом, при объединении этих двух моделей получается инструмент для поиска наиболее вероятной последовательности слов (Mohri, Pereira & Riley, 2002).

Как и для любой задачи в сфере машинного обучения (ML), для ASR большое значение имеют данные – независимые признаки, которые выделяются из аудиофайла. Для задачи автоматического распознавания речи самым популярных подходом к извлечению признаков является метод MFCC: для каждого отрезка аудио семпла длиной 25 миллисекунд выделяются 39 характеристик.

Скрытые Марковские модели для ASR состоят из скрытых переменных и признаков MFCC. Как и в обычной цепи Маркова, текущее состояние модели предсказывается предыдущим состоянием, но элементы называются «скрытыми», так как не все состояния могут быть наблюдаемы.

При условии, что Скрытая Марковская модель уже обучена, можно использовать алгоритм прямого хода, чтобы рассчитать вероятности наблюдений. Основная цель этого процесса – суммировать вероятности всех наблюдений для всех возможных состояний. Этот шаг, однако, требует достаточно искусного подхода, так как одновременное сложение последовательностей всех возможных состояний – это слишком сложный процесс – с

каждый шагом сложность структуры увеличивается в геометрической прогрессии.

Решение проблемы разбивается на несколько промежуточных этапов, а вычисления проводятся рекурсивно. В СММ проблема решается в промежуток времени *t* с помощью результата от времени *t-1* и/или *t+1*. Таким образом, даже несмотря на то, что количество последовательностей состояний экспоненциально увеличивается, рекуррентный подход к вычислению позволяет выражать их линейно.

Затем, с помощью последовательности наблюдений, находятся внутренние состояние, которые в ASR представляют собой звуки. Так автоматическое распознавание речи можно рассматривать нахождение этих внутренних состояний в аудиофайле. Этот процесс называется расшифровкой. Компоненты также вычисляются рекуррентно, и фактически, этот алгоритм повторяет алгоритм прямого хода, однако вместо сложения используется функция максимума – среди всех возможных последовательностей состояний выбирается самый вероятный путь (Chavan & Sable, 2013).

Следующий, самый сложный этап – обучение СММ-модели, за которым стоит алгоритм прямого и обратного хода. Обучение вероятностям эмиссии и вероятностям перехода – непростая задача; это проблема, где трудно отделить причину от следствия: обе вероятности очень сложно вычислить, однако их можно найти, если известна дистрибуция состояний в отрезке времени *t*, при этом сама дистрибуция состояний также вычисляется при известных вероятностях эмиссии и перехода. В машинном обучении эта проблема обычно решается с помощью классического ЕМ-алгоритма, но в автоматическом распознавании речи грамотнее использовать один из его вариантов – алгоритм прямого и обратного хода; он решает проблему в каждом шаге цикла, оптимизируя одну скрытую переменную, в то же время регулируя остальные. Следовательно, решение становится лучше в каждом шаге цикла (Chavan et al., 2013).

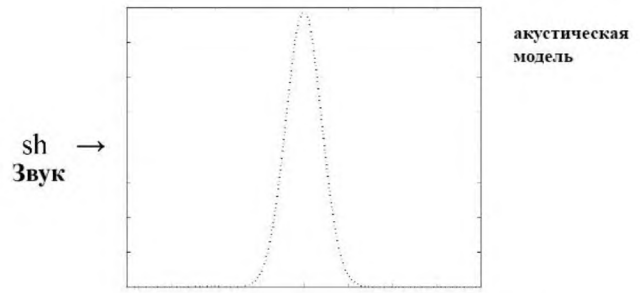


Рис.1 Нормальное распределение звука

Для реализации акустического распределения в ASR также используются таблицы транскрипций, с помощью которых создаются акустические модели для звуков из данной последовательности. На сегодняшний день технологии позволяют машинам читать гласные и согласные звуки напрямую со спектрограммы. С помощью MFCC из отрезка аудио выделяются 39 признаков (Ganchev, Fakotakis & Kokkinakis, 2005). Значение признаков звуков, как эталонных, так и выделенных из аудио-фрейма можно представить в виде нормального распределения. Для работы с характеристиками MFCC используется концепция многомерного нормального распределения. Вероятность p(x|q) выделенной характеристики x рассчитывается по её расположению относительно нормального распределения q. Для правильной работы Скрытых Марковских моделей в системе автоматической обработки речи при помощи ряда примеров звуков речи рассчитывается соответствующая им плотность распределения вероятности, которая затем классифицируется в один звук с высочайшим значением вероятности – из этого складываются данные, используемые в CMM как вероятности эмиссии.

**Метод глубокого обучения**

В настоящее время в сфере автоматического распознавания речи также используются новые подходы машинного обучения (ML), отличные от статистического метода CMM. Популярность среди разработчиков набирает метод глубокого обучения (DL), который уже широко используется для решения прикладных задач

(Deng & Liu, 2018). Глубокое обучение состоит из множества алгоритмов машинного обучения, принимающих данные в виде многослойных моделей. Эти модели – нейронные сети (НС), включающие в себя разные уровни нелинейных операций (Weng, & Yu, 2019). Алгоритмы ML обучаются глубокими нейросетями, при этом выделяется нужная информация – характерные признаки.

ASR система, полностью основанная на технологии глубокого обучения, способна напрямую преобразовывать последовательность аудио данных в последовательности слов. Структура такого механизма состоит из следующих частей: устройство кодирования, преобразовывающее последовательность входных речевых данных в последовательность признаков; блок выравнивания, совмещающий последовательность признаков с языковыми данными; декодер, который расшифровывает последовательности в выходные данные. Однако именно в DL автоматическом распознавании речи такая структура встречается редко, так как в подобных системах зачастую нет деления на модули. Но такая модель даёт представление о том, как работают нейронные сети в ASR, а также проводит аналогию со статистическими системами, состоящими, как правило, из множества модулей. СММ-модули автоматического распознавания речи могу быть замещены одной глубокой сетью, причем это касается как отдельных частей системы, так и всех механизмов в целом.

Искусственные нейронные сети рассматриваются как замена СММ в прикладных задачах уже на протяжении двух десятилетий, так как они, в отличии от Скрытых Марковских моделей, лишены такого недостатка, как статистическая неэффективность при моделировании данных, расположенных на или около нелинейного множества в пространстве данных.

Ещё в начале 1990-ых исследователи смогли добиться успеха в использовании нейронной сети с одним слоем в задаче предсказания состояний Скрытой Марковской модели для акустического моделирования. Однако в то время ни оборудование, ни линейные алгоритмы не были пригодны для обучения нейронных сетей с несколькими скрытыми слоями на больших объемах данных, а потому практической пользы от использования НС в автоматическом распознавании речи не было.

За последние годы прогресс в сфере алгоритмов машинного обучения и создания компьютерного оборудования послужили толчком к появлению более эффективных способов обучения глубоких нейронных сетей, содержащих множество слоёв нелинейных скрытых элементов и огромных слой с выходными данными, нужный для работы с огромным количеством состояний СММ, которые возникают при моделировании каждого звука определенным количеством различных триграммных Скрытых Марковских моделей, учитывающих звуки обеих сторон. С помощью новых методов обучения сразу несколько исследовательских групп по всему миру смогли построить как акустические, так и языковые нейросетевые модели, который способны превосходить в эффективности СММ-модели для автоматического распознавания речи; в исследованиях использовалось большое разнообразие наборов данных, включая огромные массивы лексических данных (Hinton G. et al., 2012).

Обучение языковых и акустических НС-моделей обычно происходит в два этапа: в начале определяющим признаки слоям по очереди присваиваются начальные установки с помощью подбора стеков генеративных моделей, каждая из которых имеет один слой скрытый переменных – эти генеративные модели обучаются без использования какой-либо информации о состояниях Скрытых Марковских моделей, которые акустическая или языковая модель должна будет выделить; второй этап предполагает использование каждой генеративной модели в стеке для присваивания начальных установок одному слою скрытых элементов в глубокой нейронной сети, после чего вся сеть дифференциально настраивается для предсказания искомых состояний Скрытых Марковских моделей. Эти состояния выделяются при помощи базовой статистической системы для того, чтобы произвести динамическое, искусственное выравнивание.

**Виды нейронных сетей**

В сфере автоматического распознавания речи используются нейронные сети нескольких видов.

Нейронные сети прямого распространения – модели, имеющие определенную структуру и количество входных и выходных узлов – это набор сенсоров, которые собирают сигналы и мгновенно

превращают их в выходные данные. Обученная нейронная сеть такого типа всегда будет выводить один и тот же результат при вводе. Иначе говоря, это не зависящая от времени модель. К этому типу относятся свёрточные нейронные сети (CNN) – они являются полезным инструментом в задачах распознавания общих моделей сигналов. В автоматическом распознавании речи они могут использоваться на этапе предобработки, а также при акустическом и языковом моделировании в тандеме со Скрытыми Марковскими моделями. В целом, свёрточные нейронные сети слабо эффективны в ASR, и используется в распознавании речи только для вторичных задач.

Проблемы автоматического распознавания речи по своему характеру существенно отличаются от задач, которые приходится решать свёрточной нейронной сети. Во-первых, в ASR невозможно выстроить набор входных узлов в соответствии с действительной длинной последовательности входных данных – аудио файл может охватывать миллионы единиц данных. В то же время, контекст определенной части входных данных с большей вероятностью связан с входными данными, обрабатываемыми до и после него. Естественно, для решения этих проблем используют уже готовые инструменты – дополняются существующие технологии нейронных сетей: в модели вводится *время*.

Такие нейронные сети называются рекуррентными (RNN). Выходные данные для определенного набора входных данных в них основывается также на предыдущем вводе. Конечно, такая структура тоже имеет свои недостатки, и создаёт много новых проблем. Например, не очевидно, насколько сильно в прошлое должна смотреть система, генерирую выходные данные. Нейронные сети долгой краткосрочной памяти пытаются решить эту проблему – они вводят специальные узлы, которые смешивают последний ввод с более ранними входными данными.

Как правило, все нейронные сети с временной зависимостью, то есть те, что обрабатывают предыдущие вводы для генерирования выходных данных, считаются лучшими НС для работы с речью и текстом, но рекуррентные нейронные сети выделяют как крупнейший класс таких моделей.

Нейронные сети эффективны в задачах моделирования нелинейных и почти нелинейных функций, однако они работают

корректно только при обработке коротких по времени сигналов; если дело касается непрерывной естественной речи, то нейросети почти не справляются с поставленной задачей – DL неспособно моделировать временные зависимости для продолжительных сигналов (Audhkhasi et al., 2018). Но, несмотря на это, нейронные сети широко используются как инструмент предобработки для основанных на CMM систем распознавания речи (Abdel-Hamid et al., 2014).

К преимуществам DL над CMM можно отнести то, что нейросетевой компонент ASR всегда является единой системой: в статистическом распознавании речи разные модули приложения используют разные методы обучения, а значит все компоненты требуют отдельной независимой оптимизации. Обучения такой системы – процесс очень сложный и запутанный. Нейронные сети не обладают этим недостатком из-за своей цельности – даже гибридные системы оптимизируются легче при наличии в них компонентов DL из-за того, что для их обучения требуется лишь одна функция, подходящая под критерий окончательной оценки – требуемый общей оптимизацией показатель.

Кроме того, основанные на CMM модели требуют вторичной обработки выходных данных, так как результаты ASR имеют особое представление внутри системы. DL же в свою очередь позволяет последовательности входных данных совмещаться напрямую с выходными данными, текстом.

Нейронные сети, однако, проигрывают статистическим моделям в работе с лексическими данными. CMM имеют дополнительные языковые модели, цель которых – создавать для системы наборы лексических данных. В моделях глубокого обучения эта лексика задана лишь обучающими данными – транскрибированными вручную человеком текстами, из-за чего охват модели очень ограничен, что ведет к проблемам в работе с лингвистически разнородной выборкой.

К минусам нейронный сетей в ASR также можно отнести то, что они, в отличии от CMM, требует разработки почти с нуля для большинства сервисов – компании не тратят время на создание новых алгоритмов распознавания речи только ради внедрения НС; улучшение уже имеющихся механизмов способно показать объективно лучший результат, кроме того, это не требует

оптимизации всей системы к новым алгоритмам – обновление одного компонента не влияет на структуру (Wang, Wang & Lv, 2019).

Для сравнения эффективности систем распознавания речи в сфере ASR существует несколько разных стандартов измерения точности, одним из которых является WER (word error rate, частота ошибочных слов) – общий стандарт измерения точности работы систем распознавания речи и автоматического перевода. В основе этого стандарта лежит редакционное расстояние Левенштейна, работающее на словесном уровне. Концепция WER довольна проста и интуитивна: неправильно распознанные слово делятся на правильные слова.

Внутри же стандарта измерения ошибочных слов процент неверного результата рассчитывается по количеству слов, которые были удалены (deleted), заменены (substituted) или добавлены (inserted) в процессе обработки речи. Алгоритм делит сумму этих слов на количество слов, которые должны присутствовать в тексте. Для определения правильности обработанного текста требуется эталонная расшифровка аудиозаписи, произведённая человеком вручную.

$$\text{WER} = \frac{\text{удаление} + \text{замена} + \text{добавление}}{\text{количество слов в ручной расшифровке}}$$

*Формула 1*

WER имеет несколько недостатков: он не работает с пунктуацией, с особенностями речи говорящего; ошибки, связанные со служебными словами, он обрабатывает наравне со знаменательными; жёстко определяются любые ошибки в тексте (полностью неверное слово обрабатывается также, как и слово с одним неверным символом). Но несмотря на это, WER считается ценным и очень полезным инструментом сравнения разных систем обработки речи (Wang, Acero & Chelba, 2003).

**Эксперимент**

Системы ASR состоят из двух относительно независимых компонентов – акустической и языковой модели; части системы нередко используют для работы разные технологии. Для

сравнения эффективности НС и СММ в автоматическом распознавании речи следует анализировать выборку, состоящую из сервисов, использующих как Скрытые Марковские модели, так и нейронные сети в акустической и языковой моделях. Сравнивая между собой сервисы с одинаковым механизмом работы в одной и той же модели, а затем сравнив их с сервисами, где та же модель основывается на другом методе, можно сделать выводы об успешности работы алгоритмов.

Таким образом, минимальное количество требуемых для анализа систем – 4: половина из них имеет в основе своих языковых моделей НС, другая – СММ; то же самое и с акустической моделью.

| Акустическое моделирование / Языковое моделирование | Нейронные сети | Скрытые Марковские модели |
|---|---|---|
| Нейронные модели | Google Cloud Speech API | Microsoft Azure: Speech API |
| Скрытые Марковские модели | Amazon Transcribe AWS | IBM Watson Speech To Text |

*Таблица 1*

Лидеры области нередко предоставляют разработчикам открытый доступ к программному интерфейсу своих приложений (API); его удобно использовать в задаче сравнения точности работы разных систем распознавания речи.

Приложение Google Cloud Speech API от компании Google позволяет разработчикам использовать простой API для перевода живой речи в текст на 120 языках и диалектах. Все компоненты ASR от Google работают на основе нейронный сетей глубокого обучения, в частности, рекуррентных нейронных сетях долгой краткосрочной памяти (LSTM RHC).

LSTM RHC также используются в приложении для автоматического распознавания речи от Amazon при акустическом моделировании. Transcribe AWS поддерживает 24 языка и диалекта, а особенностью сервиса является простое и интуитивно понятное добавление собственного словаря. Языковое же

моделирование система выполняет с помощью СММ.

Скрытые Марковские модели совместно со свёрточными нейронными сетями составляют основу акустического моделирования в ASR-сервисе от Microsoft (Kĕpuska & Bohouta, 2017). В Azure Speech API языковое моделирование представлено только свёрточными НС. ASR-система Microsoft распознает 8 языков и диалектов.

Приложение Watson Speech To Text, в свою очередь, основывается полностью на традиционном статистическом методе. И акустическая модель, и языковая работают с помощью технологии Скрытых Марковских моделей. Сервис от IBM поддерживает 14 языков или диалектов.

Для сравнения систем автоматического распознавания речи их алгоритмы тестируются на аудио семплах живой речи. При разработке ASR разработчики зачатую пользуются крупными языковыми медиа корпусами – в них представлены данные, подобные тем, с которыми системе приходится работать в её обычном сценарии использования; например, самый известный набор данных для тестирования ASR, корпус Switchboard, включает в себя примерно 260 часов аудио записанных телефонных разговоров.

Однако для сравнения эффективности работы компонентов ASR-системы, гораздо практичнее использовать специально созданные наборы данных меньшего объёма, удовлетворяющие требуемым критериям данного исследования.

Для эксперимента были отобраны 14 аудио файлов продолжительностью от 1 до 10 минут. В выборке присутствуют примеры живой естественной речи разных диалектов и акцентов английского, записанные в разных условиях: от вырезок из новостных блоков до аудиосообщений, записанных с помощью мобильного устройства. Аудио семплы были извлечены из видео на YouTube (все ролики имеют лицензию Creative Commons) и конвертированы в формат WAV, так как сервисы ASR хуже работают со сжатым аудио.

Каждому аудио файлу был присвоен индекс – короткое имя, более практичное для использования в эксперименте:

“afr” представляет пример естественной речи на африканском варианте английского языка говорящей-женщины. В семпле

присутствует музыкальный фон и шумы, на заднем плане также различимы голоса других говорящих. Продолжительность аудио – 6 минут 5 секунд.

Файлы с индексом “am” – примеры американского разговорного английского:

“am_news” – отрывок из новостного блока; в файле присутствует речь нескольких говорящих, музыкальный фон, различные шумы. Продолжительность аудио – 9 минут 25 секунд.

“am_rus” – файл с примером речи мужчины, говорящего на американском английском, который имеет акцент, характерный для носителей восточнославянского языка. На заднем плане присутствуют шумы. Продолжительность аудио – 3 минуты 27 секунд.

“am_talk” – отрывок диалога из американского телевизионного шоу. Говорящие – мужчина и женщина, носители разных диалектов американского английского. На фоне присутствуют незначительные шумы. Продолжительность аудио – 3 минуты 26 секунд.

“am_rl” – пример живого монолога носительницы калифорнийского варианта американского английского. В файле практически отсутствуют какие-либо шумы. Продолжительность аудио – 2 минуты 40 секунд.

Индекс “br” обозначает примеры речи на британском варианте английского:

“br_news” – отрывок из британского новостного шоу – нарезка из интервью местных политиков. В аудио представлена речь нескольких говорящих; на заднем плане присутствуют различные шумы и музыкальный фон. Продолжительность аудио – 2 минуты 55 секунд.

“br_podcast” – отрывок диалога носителей мужского пола из шоу-подкаста. В файле практически отсутствуют какие-либо шумы. Продолжительность аудио – 2 минуты 51 секунда.

“br_rl” – живой монолог носительницы суссекского диалекта английского языка. На фоне присутствуют незначительные шумы. Продолжительность аудио – 1 минута 32 секунды.

“br_voicemess” – голосовое сообщение, монолог, записанный мужчиной-носителем британского английского. В файле практически отсутствуют какие-либо шумы, но аудио отличается

низким качеством записи. Продолжительность семпла – 1 минута 22 секунды.

“hk” – представляет собой запись голоса мужчины, говорящего на гонконгском варианте английского. На фоне присутствуют незначительные шумы. Продолжительность аудио – 2 минуты 47 секунд.

Семплы с индексом “ind” – примеры речи англоговорящих, имеющих акцент, характерный для жителей Индии и Пакистана:

“ind_br” – монолог носителя британского английского мужского пола, обладающего индийским акцентом. В файле присутствуют незначительные шумы, но аудио отличается низким качеством записи. Продолжительность семпла – 1 минута 27 секунд.

“ind_news” – отрывок из индийского телевизионного шоу. В семпле присутствует речь нескольких говорящих, а также музыкальный фон и различные шумы. Качество записи – низкое. Продолжительность аудио – 2 минуты 24 секунды.

“mpl” – файл, который представляет собой интервью с несколькими носителями разных вариантов и акцентов английского языка. В записи присутствует музыкальный фон и различные шумы. Продолжительность аудио – 5 минут 32 секунды.

Для каждого элемента набора данных был создан файл формата TXT с транскрибированным вручную аудио. Эти тексты принимаются за эталонные данные (Ground Truth) – с ними сравниваются результаты работы ASR-систем. Для тестирования важно иметь максимально похожие данные, поэтому при транскрибировании текст был нормализован с учётом особенностей наших систем – в них отсутствуют пунктуационные знаки, прописные буквы и числа (как цифры, так и слова).

Для работы с API приложений для автоматического распознавания речи аудио семплы из выборки обрабатываются с помощью программы на языке Python.

*Google Cloud Speech API*

Доступ к API Google Cloud Speech условно бесплатный – идентификация пользовательского тарифа осуществляется с помощью специального ключа. На сайте проекта представлена вся нужная документация о сервисе, а также подробная инструкция

для разработчиков.

Google предлагает три вида ASR: потоковое распознавание, распознавание длинных файлов (через специальное облачное хранилище Google Cloud) и распознавание коротких файлов (до 60 секунд). Для исследования мы можем воспользоваться последним вариантом обрабатывая семплы по частям.

Обработаем наши аудио семплы с помощью программы на Python (Google также предлагает ещё 9 языков программирования). Наш файл должен быть транскрибирован, и результат ASR записан в файл формата TXT с тем же индексом, что и оригинальный, но с префиксом “gc”.

*IBM Watson Speech To Text*

В ASR-сервисе от IBM распознавание длинных аудио доступно локально, поэтому программа для Watson Speech To Text работает с целыми семплами. Компания также предоставляет всю документацию и инструкцию для системы на своём сайте. IBM Watson Speech To Text – это условно бесплатный проект; пользователь идентифицируется по индивидуальному ключу. Как и предыдущая, данная программа записывает выходные данные в новый файл с индексом оригинального текста, но с добавленным префиксом “ibm”.

*Microsoft Azure: Speech API*

Как и предыдущие сервисы, Azure: Speech API – условно бесплатный. Компания также предлагает потоковое распознавание, и распознавание больших и малых файлов, среди которых мы снова выбираем последнее. Для идентификации пользователя Microsoft использует личный код. Вывод данных происходит в файлы с префиксом “ms”.

*Amazon Transcribe AWS*

В сервисе распознавания речи от Amazon есть функция обработки аудио файлов с помощью виртуальной машины. Наш набор данных загружается в облачное хранилище Amazon, где мы и работаем с ним. Готовые выходные данные мы получаем в формате JSON – конвертируем их в TXT файлы с префиксом “amazon”.

Не все представленные в выборке приложения имеют опцию выбора диалекта языка для распознавания, поэтому для чистоты эксперимента используется либо распознавание с помощью

стандартной опции для английского языка, либо американский вариант английского – выбор этой опцией мотивирован тем, что, для работы на рынке стран Северной Америки, ASR-системы должны обладать возможностью распознавать более широкий диапазон фонологических конструкций; это связано с тем, что в Канаде и США присутствует большое количество различных акцентов и диалектов.

Результаты распознавания транскрибируются тем же способом, что и эталонные тексты. Как уже упоминалось, все ASR-сервисы работают по-разному: одни распознают числа как слова, другие – как цифры; некоторые системы умеют распознавать интонацию и расставлять пунктуационные знаки, остальные – нет. Все тексты проверяются и редактируются вручную – данные приводятся к максимально похожему виду, при этом в них не изменяются сами транскрибированные слова – принципиально важно сохранить ошибки, которые сделали ASR-сервисы (Zechner & Waibel, 2000).

С помощью импортированного модуля WER в Python транскрибированные тексты сравниваются с оригинальными данными. Полученные результаты вносятся в объединённую таблицу, где также рассчитывается среднее арифметическое значений – оценка общей точности систем. Для удобства результаты тестирования представляются в процентах, а дроби сокращаются до сотых.

**Результаты и обсуждение**

По итогам тестирования систем ASR можно судить о том, какой из методов автоматического распознавания речи является более эффективным. Общая оценка точности говорит о том, что приложения от Amazon и IBM лучше остальных распознают живую речь; хуже всех – сервис от Microsoft. То есть системы с языковыми моделями, основанными на CMM, являются наиболее эффективными средствами ASR. Google Cloud Speech API показывает удовлетворительные результаты, из чего можно сделать вывод о том, что глубокое обучение также полезно для языковой и акустической моделей, а проблема Microsoft Azure: Speech API заключается в плохой оптимизации акустической части приложения либо её совместной работы с языковой моделью.

| | amazon | ibm | gc | ms |
|---|---|---|---|---|
| **afr** | 37,56% | 49,39% | 53,26% | 45,53% |
| **am_news** | 7,76% | 21,34% | 16,75% | 31,21% |
| **am_rus** | 21,83% | 26,42% | 30,45% | 36,33% |
| **am_talk** | 29,92% | 37,95% | 57,12% | 37,69% |
| **am_rl** | 24,52% | 33,25% | 48,34% | 100% |
| **aus** | 36,53% | 56,62% | 68,37% | 42,52% |
| **br_news** | 24,74% | 41,23% | 51,54% | 67,35% |
| **br_podcast** | 42,11% | 55,63% | 69,36% | 61,93% |
| **br_rl** | 27,84% | 51,89% | 44,30% | 40,50% |
| **mpl** | 21,35% | 26,82% | 36,58% | 46,60% |
| **hk** | 16,80% | 18,13% | 20,53% | 43,46% |
| **ind_br** | 18,11% | 25,66% | 32,07% | 27,16% |
| **ind_news** | 39,57% | 50,63% | 44,25% | 80,42% |
| **br_voice mess** | 10,66% | 11,33% | 13,33% | 44,66% |
| **Общая оценка** | **25,66%** | **36,16%** | **41,88 %** | **50,38 %** |

*Таблица 2*

Нейросетевая акустическая модель показывает себя лучше в распознавании разных диалектов и акцентов языка; Скрытые Марковские модели проигрывают глубокому обучению в акустическом моделировании. Хуже всех системы справились с африканским и британскими вариантами английского, однако, стоит заметить, что с распознаванием британского английского неплохо справился сервис от Amazon – это говорит о том, что обучение моделей происходило на ограниченной в диалектном разнообразии выборке данных.

Кроме того, акустическому моделированию на основе нейронных сетей лучше даётся распознавание нечёткой и плохо различимой речи с музыкой и шумом на фоне: это можно наблюдать при сравнении результатов распознавания семплов “afr”, “am_news”, “br_rl”, “br_podcast”, “ind_news”.

Самый лучший результат распознавания данных принадлежит Amazon Transcribe AWS в файле “am_news”. Наихудшее

распознавание показал Microsoft Azure: Speech API в “am_rl” – индекс WER показывает стопроцентную неточность. Очевидно, что данный результат – ошибка измерения, так как при ручном анализе документов можно видеть, что тексты частично совпадают; причина такого просчёта скорее всего кроется в небезупречном алгоритме WER и в слабой акустической обработке сервиса Microsoft – приложения не смогло распознать целые реплики аудио файла.

**Выводы**

Анализ полученных результатов сравнения показал, что самой эффективной является система, у которой акустическое моделирование основано на работе нейронных сетей, а языковое – на CMM. Нейросетевые языковые модели в целом оказались менее эффективны чем статистические. Напротив же, акустические модели, основанные на механизмах Скрытых Марковских моделей, справляются с задачей распознавания речи хуже нейросетевых аналогов – в особенности это касается примеров нестандартной речи.

В целом, можно сделать вывод о том, что точность работы системы автоматического распознавания речи очень сильно зависит от оптимизации работы её элементов – как по отдельности, так и вместе. Отсутствие оптимизации в структуре автоматического распознавания речи приводит к тому, что результаты работы приложения могут показывать стопроцентную неточность распознавания при жестком сравнении стандартом WER несмотря на то, что акустическое и языковое моделирования система выполняет успешно. И глубокое обучение, и Скрытые Марковские модели, и гибридные методы – все они эффективны только при условии, если приложение было правильно обучено; здесь также важную играет роль набор обучающих данных. Сами по себе эти инструменты как таковые невозможно объективно сравнить между собой.

Сравнительный анализ систем распознавания речи также показывает, что на данном этапе развития нейронные сети (в частности, технология глубокого обучения) сами по себе не являются более эффективным подходом к лингвистическому и акустическому моделированию в ASR чем Скрытые Марковские модели. Оба метода показывают высокую точность распознавания

речи при правильной оптимизации системы и её частей. Глубокое обучение показывает лучшую производительность в определённых случаях в то время, как Скрытые Марковские модели также показывают удовлетворительный результат при этих же условиях и отличный при других.

В конечном итоге – только разработчикам решать, какой механизм должен лежать в основе акустического и языкового моделирования ASR-системы – нейросетевые или статистические машины, лишь в их руках сделать их максимально точными. Однако стоит отметить, что технологии НС сегодня развиваются очень стремительно, поэтому перевод моделей ASR на новый алгоритм без увеличения точности распознавания может стать полезной инвестицией в будущее развитие системы.

Литература

1. Abdel-Hamid O. et al. Convolutional neural networks for speech recognition // IEEE/ACM Transactions on audio, speech, and language processing. – 2014. –Vol. 22. – №. 10. – Pp. 1533-1545.
2. Audhkhasi K. et al. Building competitive direct acoustics-to-word models for English conversational speech recognition // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2018. – С. 4759-4763.
3. Audhkhasi K. et al. Direct acoustics-to-word models for english conversational speech recognition // arXiv preprint arXiv:1703.07754. – 2017.
4. Chavan R. S., Sable G. S. An overview of speech recognition using HMM // International Journal of Computer Science and Mobile Computing. – 2013. – Vol. 2. – №. 6. – Pp. 233-238.
5. Deng, L., Liu Y. (Ed.). Deep Learning in Natural Language Processing. – Springer, 2018.
6. Ganchev T., Fakotakis N., Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task // Proceedings of the SPECOM. – 2005. – Vol. 1. – №. 2005. – Pp. 191-194.
7. Hinton G. et al. Deep neural networks for acoustic modeling in speech recognition // IEEE Signal processing magazine. – 2012. – Т. 29.

8. Juang, B. H., Rabiner L. R. Automatic speech recognition–a brief history of the technology development. – Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara. – 2005. – Vol. 1. – Pp. 67.

9. Këpuska, V., Bohouta G. Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx) //Int. J. Eng. Res. Appl. – 2017. – Vol. 7. – №. 03. – Pp. 20-24

10. Mohri, M., Pereira F., Riley M. Weighted finite-state transducers in speech recognition // Computer Speech & Language. – 2002. – Vol. 16. – №. 1. – Pp. 69-88.

11. Wang, Y. Y., Acero A., Chelba C. Is word error rate a good indicator for spoken language understanding accuracy // 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721). – IEEE, 2003. – Pp. 577-582.

12. Wang D., Wang X., Lv S. An Overview of End-to-End Automatic Speech Recognition // Symmetry. – 2019. – Vol. 11. – №. 8. – Pp. 1018.

13. Weng, C., Yu D. A Comparison of Lattice-free Discriminative Training Criteria for Purely Sequence-Trained Neural Network Acoustic Models // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2019. – Pp. 6430-6434.

14. Zechner, K., Waibel A. Minimizing word error rate in textual summaries of spoken language // 1st Meeting of the North American Chapter of the Association for Computational Linguistics. – 2000.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing, 22*(10), 1533-1545.

Audhkhasi, K., Kingsbury, B., Ramabhadran, B., Saon, G., & Picheny, M. (2018, April). Building competitive direct acoustics-to-word models for english conversational speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4759-4763). IEEE.

Audhkhasi, K., Ramabhadran, B., Saon, G., Picheny, M., & Nahamoo, D. (2017). Direct acoustics-to-word models for English

conversational speech recognition. arXiv preprint arXiv:1703.07754.

Chavan, R. S., & Sable, G. S. (2013). An overview of speech recognition using HMM. *International Journal of Computer Science and Mobile Computing*, *2*(6), 233-238.

Deng, L., & Liu, Y. (Eds.). (2018). *Deep Learning in Natural Language Processing.* Springer.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005, October). Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proceedings of the SPECOM* (Vol. 1, No. 2005, pp. 191-194).

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., ... & Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine, 29.*

Juang, B. H., & Rabiner, L. R. (2005). *Automatic speech recognition – a brief history of the technology development.* Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara.

Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Int. J. Eng. Res. Appl, 7*(03), 20-24.

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language, 16*(1), 69-88.

Wang, Y. Y., Acero, A., & Chelba, C. (2003, November). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)* (pp. 577-582). IEEE.

Wang, D., Wang, X., & Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. *Symmetry*, *11*(8), 1018.

Weng, C., & Yu, D. (2019, May). A Comparison of Lattice-free Discriminative Training Criteria for Purely Sequence-Trained Neural Network Acoustic Models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6430-6434). IEEE.

Zechner, K., & Waibel, A. (2000). Minimizing word error rate in textual summaries of spoken language. In *1st Meeting of the North American Chapter of the Association for Computational*

**УДК 811.112.2**
**https://doi.org/10.25076/vpl.36.02**
**Н.А. Воробьева**
**П.Н. Хроменков**
**Московский государственный областной университет**

# ЛИНГВОПРАГМАТИКА НЕМЕЦКИХ МУЗЫКАЛЬНО-ПОЭТИЧЕСКИХ ТЕКСТОВ

*В статье анализируется влияние музыкально-поэтических текстов на слушателей. Любой поэтический текст может восприниматься по-разному, а в соединении с музыкой влияние может усилиться. На это влияет множество факторов, зависящих не только от адресанта, но и от реципиента, а также других окружающих факторов. Данная работа основывается на значимости роли языковой составляющей в процессе создания и восприятия текстов песен современного музыкального исполнителя музыкального стиля «метал-музыка», являющегося на данный момент одним из инструментов воздействия, как на коммуникативную активность человека, так и на его восприятие. Предметом данного исследования являются тексты песен всемирно известной немецкой метал-группы Rammstein. В задачи данного исследования входят определение понятий «поэтического и музыкального дискурса», «метал-музыки»; идейный смысл текстов Rammstein; особенности восприятия, интерпретации и роль ритма в музыкально-поэтическом сообщении; анализ лингвистических особенностей музыкально-поэтического текста на примере текста одной из песен Rammstein, а также опрос на тему восприятия творчества данной группы. В ходе исследования теоретически и экспериментально было доказано, что всякий поэтический текст действительно может восприниматься совершенно по-разному, на что влияет множество факторов, зависящих не только от адресанта, но и от реципиента.*

*Ключевые слова: поэтический текст, поэтический дискурс, музыкальный дискурс, метал-музыка, немецкий язык, Rammstein, влияние, образ мышления, сознание.*