Санкт-Петербургский государственный университет

ХОЛЯВИН Павел Андреевич

Выпускная квалификационная работа

Адаптация произносительного словаря для автоматического распознавания разных типов речи

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5715. «Общая и прикладная фонетика (General and Applied Phonetics)»
Профиль «Речевые технологии»

Научный руководитель:
доцент, Кафедра фонетики и методики преподавания иностранных языков,
Кочаров Даниил Александрович

Рецензент:
старший научный сотрудник, Лаборатория речевых и многомодальных интерфейсов, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», Санкт-Петербургский институт информатики и автоматизации Российской академии наук,
Кипяткова Ирина Сергеевна

Санкт-Петербург

2021

Оглавление

Введение

Данная работа посвящена вопросам адаптации и оптимизации произносительных словарей для систем автоматического распознавания (АРР) русской речи применительно к разным её типам. Известно, что на речь человека влияет множество факторов, и это в немалой степени касается её фонетической стороны. В зависимости от речевой ситуации одни и те же лексические единицы могут реализовываться в различных формах; более того, зачастую эти формы значительно отличаются от вариантов, предложенных в орфоэпических словарях.

Системы автоматического распознавания речи чаще всего пользуются именно такими орфоэпическими транскрипциями, что не может не отражаться на качестве распознавания. Поиск транскрипций, которые бы наилучшим для системы АРР образом отражали реальное произношение, является актуальным и на сегодняшний день [Adda-Decker, Lamel, 2018; Lukeš и др., 2018], и не в меньшей степени он является актуальным для русского языка, который характеризуется разного рода фонетическими процессами: изменение гласных вследствие коартикуляционных процессов, их редукция в безударном положении, ассимилятивные изменения согласных и др.

Объектом данного исследования является автоматическое распознавание речи; предметом — методы адаптации произносительных словарей и зависимость качества распознавания от использования тех или иных словарных транскрипций.

Целью исследования является поиск способов оптимизации существующих словарей или создания новых таким образом, чтобы результирующий словарь обеспечил улучшение качества автоматического распознавания.

Для достижения данной цели были выполнены следующие задачи:

1. Обзор литературы, описывающей существующие методы и подходы к распознаванию речи, а также созданию и модификации транскрипций;

2. Обзор литературы, описывающей особенности русского языка, отражение которых в произносительном словаре могло бы повысить качество распознавания;

3. Предложение способов создания или модификации произносительного словаря;

4. Проведение экспериментов по автоматическому созданию транскрипций и автоматическому распознаванию речи;

5. Подведение итогов экспериментального исследования и оценка предложенных методов.

Результаты данного исследования могут помочь в создании новых систем АРР, адаптированных под разные типы речи, а также в создании систем автоматического выравнивания.

Работа состоит из введения, трёх глав и заключения. В первой главе освещены устройство систем распознавания речи, особенности создания словарей для задач АРР, методы автоматического создания транскрипций. Во второй главе рассмотрены фонетические характеристики и особенности русской разговорной речи, имеющие значение для поставленной задачи. В третьей главе описаны материал и методика исследования, освещены поставленные эксперименты и приведены их результаты.

Глава 1. Автоматическое распознавание речи и автоматическая обработка транскрипций

1.1. Задачи автоматического распознавания речи

Основной задачей исследований, связанных с автоматическим распознаванием речи (АРР) является разработка таких систем, которые бы сопоставляли акустическому сигналу определённую последовательность слов [Jurafsky, Martin, 2014, с. 233].

В зависимости от поставленной задачи подходы и архитектуры таких систем могут быть различными. Во-первых, системы могут различаться по количеству известных слов: оно может ограничиваться единицами или десятками (например, различение слов «да» или «нет» или распознавание цифр), а может превышать несколько десятков и сотен тысяч. С другой стороны, существуют различные системы с точки зрения характера распознаваемой ими речи, её слитности и естественности. Так, задача распознавания изолированных слов представляется гораздо более лёгкой, чем распознавание слитной речи; распознавание речи, адресованной машине (например, чтение вслух) легче, чем распознавание разговора между людьми. В-третьих, на структуру системы влияет предполагаемый канал передачи речи: распознавание речи в зашумлённом канале, несомненно, сложнее, чем распознавание звука, записанного в студийных условиях на качественный микрофон. Наконец, важным фактором является лингвистический или физиологический характер распознаваемой речи: одна и та же система может с разной долей успешности распознавать речь взрослых и детей или, например, речь носителей разных диалектов одного и того же языка в зависимости от того, как именно она была обучена [Jurafsky, Martin, 2014, с. 234]. В данной работе будут рассматриваться системы автоматического распознавания слитной речи с использованием больших словарей, т. е. словарей, объём которых исчисляется тысячами и десятками тысяч слов [Кипяткова, Карпов, 2010].

1.2. Краткая историческая справка

Системы разного характера появились не одновременно. Самые первые из них были созданы ещё в середине XX века. Так, одной из первых является система под названием AUDREY, которая была разработана в лаборатории Белл, Нью-Йорк, США [Davis, Biddulph, Balashek, 1952]. Эта система обладала ограниченным словарём и была рассчитана на распознавание отдельных цифр. Спустя 14 лет компания IBM представила свою систему Shoebox [Lama, Namburu, 2010], словарь которой состоял из 16 слов. В 1976 году система Harpy была способна распознавать до 1000 разных слов [Lowerre, Reddy, 1976]. Её новшеством являлся лучевой алгоритм поиска, который до сих пор считается основным способом поиска наиболее вероятной гипотезы распознавания. В то же время была создана система Hearsay-I, одна из первых, способных распознавать непрерывную речь [Reddy и др., 1976]. В конце 1970-х — начале 1980-х было выяснено, что скрытые Марковские модели (HMM) хорошо описывают релевантные для распознавания признаки со статистической точки зрения. Независимо друг от друга [Lee и др., 1990] эту технику применили две группы исследователей: группа под руководством Дж. К. Бейкера в университете Карнеги — Меллона [Baker, 1975] и группа под руководством Ф. Йелинека в компании IBM [Jelinek, 1976]. Системы, основанные на структуре HMM-GMM (скрытые Марковские модели, каждое состояние которых описывается линейной комбинацией нормальных распределений), используются до сих пор [Povey и др., 2011]. К 1990-м годам системы уже были способны на дикторонезависимое распознавание слитной речи [Huang и др., 1993].

В 2000-х годах начался переход на искусственные нейронные сети и гибридные подходы, которые доминируют и сейчас [Кипяткова, Карпов, 2016; Dahl и др., 2011]. Среди разновидностей нейронных сетей можно упомянуть DNN — глубокие нейронные сети [Pan и др., 2012], которые обладают

несколькими слоями между входным и выходным; RNN — рекуррентные нейронные сети [Vinyals, Ravuri, Povey, 2012], которые обладают «памятью» — т. е. способны хранить информацию о прошлых данных, что позволяет им обрабатывать временные последовательности; CNN — свёрточные нейронные сети [Palaz, Collobert, 2015], которые используют математическую операцию свёртки для обработки входных данных, что, в частности, может позволить им работать напрямую со спектрограммой. Среди новейших типов ИНС можно упомянуть трансформеры [Vaswani и др., 2017], которые используют механизм внимания для нелинейной обработки входных данных [Jain и др., 2020].

1.3. Устройство системы АРР

В АРР широко используются методы математической статистики. Процесс распознавания можно представить с этой точки зрения как поиск наиболее вероятной для данного акустического сигнала последовательности слов с опорой на две модели: акустическую и языковую [Кипяткова, Карпов, 2016, с. 80]. Вопрос, на который отвечает система распознавания речи, можно сформулировать следующим образом: каково наиболее вероятное предложение из всех возможных в языке L при условии акустического сигнала O [Jurafsky, Martin, 2014, с. 237]. Если рассматривать звуковой сигнал как некую последовательность $O = o_1, o_2, \ldots, o_t$, а предложение, в свою очередь, как последовательность слов $W = w_1, w_2, \ldots, w_n$, то формула, описывающая ответ на вопрос, будет записываться следующим образом:

$$\hat{W} = \underset{W \in L}{argmax}\, P(W|O) \ ,$$

откуда по теореме Байеса

$$\hat{W} = \underset{W \in L}{argmax} \frac{P(O|W)P(W)}{P(O)} = \underset{W \in L}{argmax}\, P(O|W)P(W) \ ,$$

т. к. для каждой гипотезы P(O) остаётся неизменным. Итак, остаётся оценить вероятности P(O|W) и P(W). Первая вероятность и описывается акустической моделью, а вторая — языковой.

Кроме этих двух моделей, система АРР включает произносительный словарь, перечисляющий все известные ей слова и их произношения, а также декодер, который и является тем математическим аппаратом, который оценивает наиболее вероятную последовательность слов с помощью акустической и языковой модели [Nilsson, 2013, с. 7].

Итак, типичная система АРР состоит из следующих компонентов:

1. Акустическая модель;
2. Языковая модель;
3. Произносительный словарь;
4. Декодер.

Схема, показывающая связи между модулями, приведена на рис. 1. Акустическая модель, языковая модель и декодер как компоненты системы будут рассмотрены ниже; произносительный словарь же будет освещён в отдельном разделе.

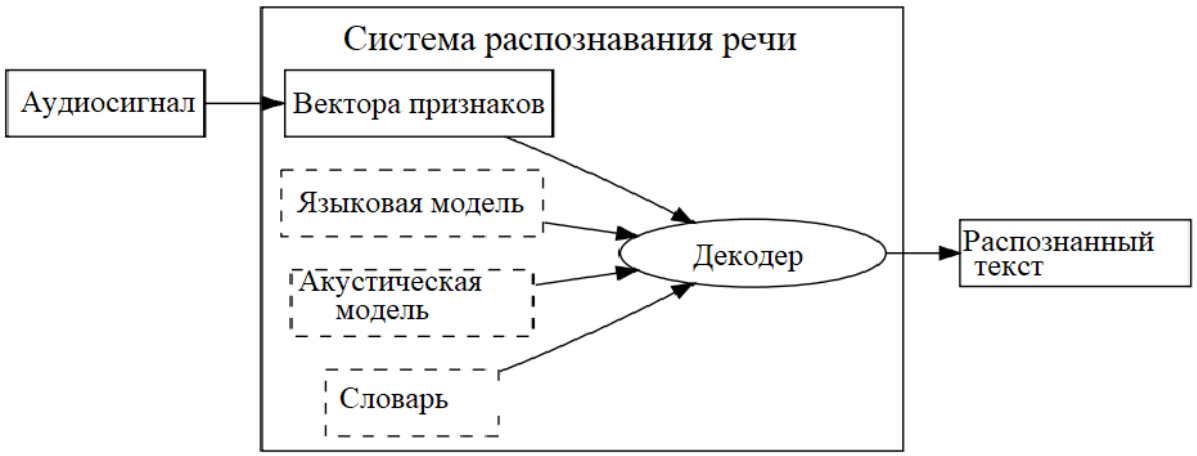

Рис. 1 - Схема типичной системы АРР

Акустическая модель (АМ) — модель, которая описывает фонетические характеристики звуков языка, для распознавания которого предназначена, а также междикторской вариативности.

Одним из самых распространённых методов создания акустических моделей является использование скрытых Марковских моделей. Скрытая Марковская модель (СММ, англ. HMM — Hidden Markov Model) — это математический аппарат, который позволяет представить последовательность некоторых событий в форме параметрического случайного процесса [Rabiner, 1989]. Впервые СММ были описаны ещё в 1972 году [Baum, 1972].

СММ представляет из себя последовательность состояний, соединённых между собой переходами. Каждому переходу ставится в соответствие вероятность перехода, т. е. вероятность выбора именно этого перехода, и функция плотности вероятности выхода, которая определяет для каждого символа из конечного алфавита вероятность получить именно его при условии данного перехода. В применении к АРР процессом, который описывает СММ, является изменение акустических признаков сигнала с течением времени. Чаще всего каждому моделируемому системой звуком ставится в соответствие СММ из трёх состояний. Пример такой СММ приведён на рис. 2.

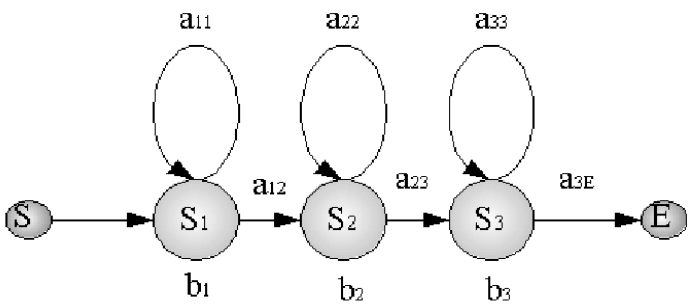

Рис. 2 — Пример СММ для одного звука из трёх состояний

СММ для единиц более высокого уровня, чем звуки (слова, предложения) принципиально не отличаются от СММ для звуков и представляют собой СММ единиц более низкого уровня, сконкатенированные между собой [Тампель, Карпов, 2016, с. 53].

Качественно другим типом акустической модели, получившим особенно широкое распространение в последнее время, является искусственная нейронная сеть [Кипяткова, Карпов, 2016]. Под искусственной нейронной сетью (ИНС, англ. ANN — artificial neural network) понимается математическая модель, которая представляет собой систему, состоящую из простых процессоров, посылающих друг другу сигналы [Haykin, 1998, с. 24]. Такая система отличается тем, что обладает способностью накапливать и использовать эмпирические данные. В процессе обучения такой сети информация хранится в коэффициентах связи между нейронами, составляющими сеть, т. е. синаптическими весами.

С математической точки зрения ИНС представляет направленный граф, вершины которого соединены связями активации и синаптическими связями [Haykin, 1998, с. 39]. ИНС можно разделить на несколько подтипов в зависимости от того, как внутри неё расположены составляющие её нейроны [Haykin, 1998, с. 43]:

1. Однослойные ИНС прямого распространения. Они состоят из слоя входных нейронов и слоя выходных (они так называются, поскольку во входном слое вычислений не производится).

2. Многослойные ИНС прямого распространения. В отличие от предыдущего типа, в них присутствует один или несколько промежуточных (т. н. скрытых) слоёв. Таким образом система способна обнаруживать сложные зависимости.

3. Рекуррентные ИНС. Отличаются от предыдущих двух типов тем, что в них присутствуют связи, направленные от нейронов в последующих слоях к предыдущим.

Для задач АРР чаще используются типы 2 и 3 [Кипяткова, Карпов, 2016, с. 83].

ИНС могут быть обучены по различным методикам, которые тем не менее подчинены определённым конвенциям. В процессе обучения на каждой его итерации нейронная сеть проводит сравнение результата на выходе с желаемым и, в зависимости от обнаруженного расхождения, исправляет синаптические веса так, чтобы результаты совпадали [Haykin, 1998, с. 73].

Стоит также упомянуть, что в АРР нередко находят применение гибридные, или же тандемные, системы, которые сочетают СММ и ИНС [Кипяткова, Карпов, 2016, с. 83]. Одной из причин популярности гибридных систем является тот факт, что СММ лучше справляются с моделированием последовательных процессов (например, последовательная смена звуков в высказывании), а ИНС — параллельных (например, разных частот в акустическом спектре) [Gevaert, Tsenov, Mladenov, 2010, с. 2].

Языковая модель (ЯМ) является представлением информации о составе слов данного языка и их сочетаемости. Языковые модели подразделяются на два основных класса: формальные и стохастические, т. е. основанные на вероятностях [Huang и др., 2001, с. 539]. В качестве формальной языковой модели можно упомянуть теорию Хомского. Формальные грамматики могут включать в себя и вероятностные данные, вычисленные на больших корпусах [Huang и др., 2001, с. 540]. Стохастические же ЯМ оценивают не возможность, а вероятность появления той или иной последовательности слов, и, как и многие другие компоненты систем АРР, широко используют аппарат математической статистики. Среди них наиболее распространены n-граммные модели, которые оценивают вероятность встретить в произвольном тексте на данном языке последовательность из n определённых слов. На практике часто используются модели с n, равным 2 (биграммные) или 3 (триграммные). Такие модели необходимо обучать на больших корпусах [Ляшевская, 2016].

Наконец, стоит отметить, что языковое моделирование возможно также и с использованием ИНС различной архитектуры [Кипяткова, Карпов, 2016, с. 90].

Задачей *декодера* является нахождение самой вероятной последовательности слов из множества всех возможных [Gales, Young, 2007]. Существует несколько различных способов ограничить количество гипотез, такие, как сохранение списка N лучших или отбрасывание гипотез, вероятность которых находится ниже порога, определяемого вероятностью наилучшей гипотезы [Тампель, Карпов, 2016, с. 112]. Все возможные слова для удобства анализа представляются как вершины некоторого графа (англ. lattice) [Gales, Young, 2007] или дерева [Jelinek, 1976, с. 540].

Обученные системы распознавания речи подвергаются проверке с целью оценки качества распознавания. Для этого обычно выделятся часть материала, которая не входит в обучающую выборку. Такая тестовая выборка предъявляется системе, после чего результат сравнивается с действительно сказанным в материале. Возможен также подход кросс-валидации, при котором обучение происходит в несколько этапов, на каждом из которых весь материал разбивается на обучающую и тестовую часть своим образом, а общий результат работы модели усредняется [Stone, 1974].

Самым часто используемым параметром для оценки качества систем АРР является WER — Word Error Rate, т. е. процент неправильно распознанных слов. Этот параметр вычисляется при помощи расстояния Левенштейна [Левенштейн, 1965], т. е. минимального количества операций (вставка, удаление или замена), которым можно из исходной строки получить новую, между оригинальным высказыванием и гипотезой, предложенной системой АРР. Т. е. WER определяется как отношение суммы всех ошибок, совершённых системой (то есть количества вставок, удалений и замен), к общему количеству слов в изначальном высказывании, умноженное на 100 %. По аналогичной формуле

могут вычисляться и другие параметры — PER (Phone Error Rate или Phoneme Error Rate; процент неправильно распознанных звуков), SER (Sentence Error Rate; процент неправильно распознанных предложений), Syllable ER (Syllable Error Rate; процент неправильно распознанных слогов) и другие; использование того или иного параметра определяется задачей исследования или типом языка [Кипяткова, Карпов, 2012].

1.4. Произносительный словарь

Важным элементом системы АРР является ее словарь. Он определяет количество слов, известных системе, и их звуковой состав. В зависимости от размера словаря системы делятся на системы с малым словарём (единицы и десятки слов), со средним словарём (сотни слов), с большим словарём (тысячи и десятки тысяч слов) и со сверхбольшим словарём (сотни тысяч и миллионы слов) [Кипяткова, Карпов, 2010]. Необходимо отметить, что количество единиц в таком словаре будет зависеть от типа языка — например, в русском языке, который является флективным, одной лексеме в словаре могут соответствовать десятки словоформ. В то же время английский язык, который является аналитическим, потребует гораздо меньшего количества словоформ на одну лексему и, как следствие, меньшего фактического объёма словаря. Для синтетических языков можно использовать словари, основанные на морфемах, а не на отдельных словах: например, такая модель предлагается для чешского языка [Byrne и др., 2000].

Чем больше словарь системы, тем большее количество слов она сможет распознать в произвольном тексте и, соответственно, тем выше будет качество распознавания. Однако увеличение объема словаря влияет также на время распознавания и, как следствие, на требуемую для работы системы вычислительную мощность, а также повышает неопределённость и, следовательно, вероятность неправильной гипотезы [Casali, Williges, Dryden, 1990, с. 112].

С данной проблемой тесно связана проблема внесловарных слов (out-of-vocabulary words; OOV). Такими словами могут быть какие-либо необщеупотребительные термины или имена собственные. Если методы обработки таких слов предусмотрены не будут, произойдёт их замена созвучными им словами из словаря, или, с большей вероятностью, последовательностью из более коротких слов, что, в свою очередь, может повлиять и на распознавание соседних слов. Одним из методов решения данной проблемы является добавление в словарь общей модели слова, в котором возможны произвольные сочетания фонем, а вероятность появления такого слова в материале будет вычисляться на больших речевых корпусах. Ещё одним методом является распознавание сначала на уровне меньших, чем слово, единиц (примером могут служить слоги или фиксированные сочетания фонем, которые также выводятся из большого речевого материала) [Тампель, Карпов, 2016, с. 117].

Каждому слову в лексиконе системы должна быть сопоставлена его транскрипция (или несколько вариантов транскрипции). Для качественного распознавания разговорной речи транскрипции должны отражать её особенности. Составление таких словарей вручную может занять много времени и требует экспертизы, поэтому применяются методы автоматического транскрибирования по правилам [Кипяткова, Карпов, 2008; Nkosi, 2012]. Кроме того, существует методика автоматического выбора наиболее подходящих транскрипций, [Sloboda, Waibel, 1996]; в таком случае необязательно составление правил.

Одной из самых значительных проблем, связанных с произносительными словарями, является существование произносительных вариантов [Strik, Cucchiarini, 1999]. Ситуация, когда такие варианты не учтены каким-либо образом в системе распознавания речи, может значительным образом сказаться на качестве распознавания. Таким образом, исследования в области поиска

методов описания такой вариативности в терминах систем АРР представляются актуальными и перспективными, и данная проблема не теряет значимости в течение последних нескольких десятилетий [Koval, Smirnova, Khitrov, 2002; Lukeš и др., 2018].

В работе [Леонтьева, Кипяткова, 2007] приведён анализ методов создания альтернативных транскрипций для моделирования вариативности в речи с помощью динамического программирования и скрытых Марковских моделей.

Ещё в конце прошлого века стало понятно, что ручное составление произносительных словарей для автоматического распознавания речи неэффективно: во-первых, ручное добавление отдельных слов не вносит заметного улучшение в общее качество распознавания; во-вторых, такие модификации могут быть подвержены ошибкам или отклонениям от реального произнесения [Sloboda, Waibel, 1996], а также зависеть от аннотатора [Kessens, 2002, с. 31]. Метод, предложенный в [Sloboda, Waibel, 1996], заключается в следующем:

1. Обучающий материал автоматически размечается на слова.

2. Создаётся матрица фонемных ошибок для базовой системы распознавания и статистическая модель, описывающая фонемный строй языка (по тем же правилам, что и языковые модели).

3. Анализируются частотные ошибки и генерируется список последовательностей слов, для которых необходимо изменение словарных транскрипций.

4. Производится пофонемное распознавание с помощью модели, созданной в пункте 2.

5. Результирующие транскрипции, отличающиеся от существующих только звуками, которые не путаются системой (на основании матрицы ошибок), добавляются в словарь.

Такой алгоритм приводит к значительному улучшению работы системы: количество ошибок снижается на 6.3 %.

Важным подвидом произносительных словарей являются словари динамические, которые подразумевают, что произносительная модель может меняться с контекстом [Fosler-Lussier, 2000]. Среди факторов, влияющих на модель, могут быть соседние слова, скорость речи, вероятности слов, содержащиеся в языковой модели, и другие.

Более подробно автоматическое создание транскрипций будет освещено в следующем разделе.

1.5. Автоматическое создание транскрипций

На сегодняшний день в мире существует более 6 500 языков [Pereltsvaig, 2020, с. 11], и одна из наиболее остро стоящих проблем, связанных с системами APP — это их адаптация всё к новым и новым языкам с минимизацией возможных затрат. Это включает в себя и создание произносительных словарей [Schlippe и др., 2012]. Кроме того, такие исследование имеют большое значения и для создания систем автоматического синтеза речи [Bruguier, Bakhtin, Sharma, 2018].

Ряд алгоритмов подразумевает составление новых транскрипций на основе орфографической записи слов с помощью автоматических транскрипторов: такой подход носит название Grapheme to Phoneme (G2P, от графемы к фонеме) [Bisani, Ney, 2008; Deri, Knight, 2016] и применяется для задач не только распознавания, но и автоматического синтеза речи. В основе алгоритмов G2P могут лежать как правила, созданные вручную, так и статистический подход: в этом случае системы автоматически обучаются на парах слово — транскрипция; эффективность будет зависеть в том числе от количества обучающего материала и его изначального качества, и фонетическая точность не будет превышать точность обучающего материала, что не позволит

решить проблему несоответствия словарных транскрипций и реального произнесения.

Разумеется, такой алгоритм будет с разной эффективностью работать для разных языков в зависимости от их систем письма и орфографии: так, он принципиально невозможен для китайского языка, который пользуется иероглифической письменностью [Lukeš и др., 2018]. С другой стороны, такие алгоритмы успешно применяются для многих младописьменных языков с высоким соответствием фонетики и орфографии.

1.5.1. Системы, работающие по правилам

Упорядоченные наборы чувствительных к контексту правил традиционно используются для описания фонологических явлений в различных фонемных и морфемных контекстах и вне поля задач APP [Kaplan, Kay, 1994]. Каждое такое правило можно описать с помощью симметричного взвешенного конечного преобразователя. Конечный преобразователь (англ. finite-state transducer, FST) — это конечный автомат, задачей которого является проведение соответствий между двумя множествами символов [Jurafsky, Martin, 2014, с. 71]. Более того, список таких правил также может быть представлен в виде конечного преобразователя. На этом принципе строятся регулярные фонологические модели языков. Разумеется, составление таких правил невозможно без участия квалифицированного специалиста.

Среди автоматических транскрипторов, созданных по правилам, можно упомянуть транскриптор для русского языка, разработанный в исследовании [Кипяткова, Карпов, 2008]. Среди особенностей данного транскриптора — отражение редукции безударных гласных и выпадения некоторых согласных в слитной речи.

Стоит упомянуть также и исследование [Евдокимова, Скрелин, Чукаева, 2017]. В особенности данного транскриптора входят: моделирование вариативности гласных и согласных фонем и их сочетаний в пределах

фонетического слова, моделирование процессов на стыке слов (орфографических и фонетических), учёт гласных вставок в консонантных кластерах, учёт влияния паузации.

1.5.2. Машинное обучение

Создание систем автоматической транскрипции возможно и с применением методов машинного обучения (МО). Алгоритмы, использующие статистические методы, обычно основываются на больших списках слов с их произношениями [Black, Lenzo, 2003].

Статистические алгоритмы G2P могут быть основаны на ряде различных принципов: скрытые Марковские модели, взвешенные конечные преобразователи, деревья принятия решений, нейронные сети [Taylor, 2005]. Многие из них находятся в открытом доступе [Novak, Minematsu, Hirose, 2016].

Отдельно рассмотрим взвешенные конечные преобразователи (weighted finite-state transducer, WFST) [Mohri, 2004]. Под WFST понимается конечный автомат, в котором каждый переход от одного состояния к другому обладает не только входной меткой, но и выходной, а также весом (т. е. числом). WFST могут применяться для описания соответствия между двумя разными последовательностями. Веса в WFST необходимы для моделирования вероятности того или иного перехода.

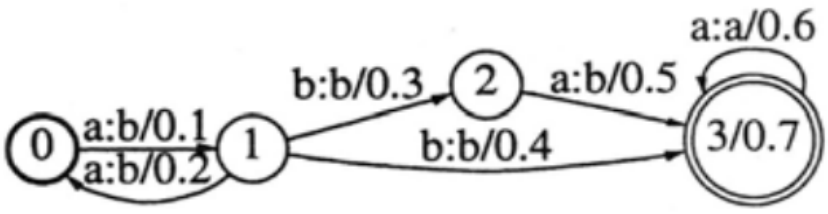

Рис. 3 — Пример простого WFST [Mohri, 2004]

В последнее время наиболее распространёнными алгоритмами являются рекуррентные нейронные сети [Bruguier и др., 2017] и особая разновидность

ИНС — Listen, Attend and Spell (LAS) [Chan и др., 2016]. Последние являются комбинированной системой для распознавания речи.

Среди недавних исследований, посвящённых автоматическому созданию транскрипций, можно упомянуть работу [Bruguier, Bakhtin, Sharma, 2018]. В ней рассматривается создание транскрипций для английского языка при помощи гибридного алгоритма, который подразумевает адаптацию ИНС, разработанной для машинного перевода. Алгоритм использует конечные преобразователи для разбиения слов на части, которым будут соответствовать одинаковые транскрипции. Затем информация о разбиениях и их соответствиях входным транскрипциям кодируется в четырёхмерный тензор. Для снижения размерности тензора используется нейросеть LSTM, результат работы которых может быть использован как вход для стандартных алгоритмов G2P — рекуррентных ИНС или LAS. Такая система способна заменить системы, основанные на ручных правилах; другим её преимуществом является её возможность применения в разных архитектурах.

Ещё одним типом нейронной сети, широко используемым в последнее время, являются трансформеры [Vaswani и др., 2017]. Их ключевой особенностью является использование механизма внимания, который позволяет им находить глобальные зависимости между данными и не полагаться на рекуррентность (которая обязывает нейросеть обрабатывать данные строго в том порядке, в котором они находятся в обучающей последовательности). Такие нейронные сети также могут использоваться для задач G2P [Yolchuyeva, Németh, Gyires-Tóth, 2020].

1.5.3. Оценка автоматически созданного словаря

Так как произносительный словарь является одним из ключевых компонентов системы APP, необходимо, чтобы в нём было как можно меньше ошибочных транскрипций. Такие ошибки могут привести не только к некачественному распознаванию, но и к ошибкам при обучении системы; в

таком случае её полный потенциал не будет реализован [Schlippe и др., 2012]. Источниками таких ошибок могут быть опечатки, использование создателями аннотации разных конвенций и транскрипционных знаков. При автоматическом создании словарей причиной ошибок также могут стать некачественные входные данные — например, ошибки в автоматическом подборе пар слово — произношение.

Существуют разные подходы к оценке автоматически созданных транскрипций и поиску ошибок в них. Так, в исследовании [Vozila и др., 2003] предлагается стохастический подход к анализу произносительного словаря и поиску вхождений, в которых вероятна ошибка. Подход основан на CMM и n-граммах. В работе [Davel, Wet, 2010] используется метод кросс-валидации при обучении системы G2P. Исследование [Giwa, Davel, 2017] предлагает метод оценки правильности словаря, который учитывал бы произносительные варианты; такой метод лучше показывает связь между ошибками словаря и качеством распознавания речи.

1.6. Обзор существующих систем APP

Поскольку распознавание речи является активно развивающимся направлением, в мире существует большое количество разнообразных систем распознавания речи. Здесь перечислены системы, находящиеся в свободном доступе и широко использующиеся исследователями.

Первой из таких широко распространенных систем стала HTK [Young и др., 2002], часто использующаяся в исследовательских целях. Данная система представляет собой инструментарий для работы со скрытыми Марковскими моделями (что позволяет применять её не только для распознавания, но и для других задач, например, синтез речи или распознавание символов). Система была изначально разработана в Кембриджском университете. В 2015 году в версию 3.5 системы была добавлена поддержка искусственных нейронных сетей [Zhang, Woodland, 2015].

Второй до недавнего времени по распространению является CMU Sphinx [Lamere и др., 2003], разработанная в университете Карнеги — Меллона. Эта система представляет собой ресурсную библиотеку для распознавания речи, содержащую утилиты для создания акустических моделей и тренировки систем распознавания, а также готовые языковые и акустические модели, а также словари. Утверждается, что эта система способна работать с показателем WER 18,8 % со словарем в 64 000 слов.

Другой системой распознавания речи является Julius [Lee, Kawahara, Shikano, 2001]. В отличие от CMU Sphinx, она представляет из себя декодер, способный работать с готовыми акустическими и языковыми моделями в виде скрытых Марковских моделей. Эта система способна работать с показателем WER 58,74 % на словаре в 20 000 слов (данные для японского языка).

Наиболее распространенной в данное время является система распознавания Kaldi [Povey и др., 2011], которая используется в данной работе. Более подробно эта система будет освещена в разделе 3.3 с точки зрения использования в данном исследовании.

1.7. Выводы по главе 1

В главе 1 были рассмотрены основные принципы, методы и задачи автоматического распознавания речи, подробно описано внутреннее устройство систем АРР и освещён ряд проблем, связанных с задачей. Кроме того, отдельно были описаны методы автоматического создания транскрипций, как с помощью правил, так и с использованием средств машинного обучения.

В данном исследовании для проведения экспериментов по распознаванию речи будет использован инструментарий Kaldi как наиболее распространённый и активно поддерживаемый из доступных; с его помощью будут создаваться акустические модели, основанные на скрытых Марковских моделях. Результаты распознавания будут оцениваться с помощью метрик Word Error Rate и Phoneme Error Rate.

Будут протестированы и сравнены методики автоматического создания транскрипций по правилам и с использованием средств машинного обучения; более подробно избранные архитектуры будут рассмотрены в соответствующем разделе.

Таким образом, глава содержит описание применённых на практике методик и технологий и тем самым представляет собой теоретическую базу для данного исследования.

Глава 2. Фонетические особенности русской спонтанной речи

В этом разделе будут освещены фонетические особенности русской речи, которые могут быть значимыми для систем АРР.

2.1. Фонетическая система русского языка

Данное исследование придерживается Ленинградской (Щербовской) фонологической школы в своих фонологических описаниях [Бондарко, 1977]. Здесь кратко описывается фонологическая система русского литературного языка в соответствии с описанием Л. В. Бондарко. В качестве транскрипционных знаков в описании будут использованы символы Международного фонетического алфавита [International Phonetic Association, 1999]. Ударения будут обозначены акутом.

В русском языке 6 гласных и 36 согласных фонем. Вокалическая система русского языка приведена в таблице 1. С точки зрения артикуляции эти гласные различаются по подъёму, по ряду и по положению губ — огубленности.

Таблица 1 — Вокалическая система русского языка

	Передний ряд	Смешанный ряд	Задний ряд
Верхний подъём	i	ɨ	u
Средний подъём	e		o
Нижний подъём			a

Русские гласные подвержены влиянию окружающего контекста. В зависимости от места образования соседних согласных форманты гласного на переходных участках изменяют свои значения. Значительные изменения наблюдаются также в сочетаниях с мягкими согласными (т. н. *i*-образный переход).

Русские гласные качественно различаются в зависимости от того, находятся ли они под ударением. Безударные гласные верхнего подъёма, как правило, менее закрытые, чем соответствующие ударные; безударные гласные нижнего подъёма, наоборот, более закрытые. Артикуляция безударных гласных также очень зависит от качества соседних согласных. В заударном слоге безударные гласные меняются сильнее всего, а абсолютное начало слова — позиция, которая провоцирует меньше всего изменений.

Консонантная система русского языка приведена в таблице 2. Согласные с точки зрения артикуляции различаются по действующему органу (по этому признаку они подразделяются на губные, переднеязычные, среднеязычные и заднеязычные; также принято выделять однофокусные и двухфокусные), по способу образования (смычные, которые подразделяются на чистые смычные и аффрикаты, щелевые и дрожащие), по участию голоса (глухие и звонкие), по твёрдости-мягкости и по участию носового резонатора (носовые и ротовые). По соотношению шума и голоса согласные делятся на шумные и сонорные (сонанты).

Таблица 2 — Консонантная система русского языка

		Губные	Переднеяз.		Среднеяз.	Заднеяз.
			Однофок.	Двухфок.		
Смычные	Шумные	p pʲ b bʲ	t tʲ t͡s d dʲ	t͡ʃ		k kʲ g gʲ
	Сонанты	m mʲ	n nʲ			
Щелевые	Шумные	f fʲ v vʲ	s sʲ z zʲ	ʃ ʃʲː ʒ		x xʲ
	Сонанты		l lʲ		j	
Дрожащие			r rʲ			

2.2. Междикторская и внутридикторская вариативность

Носители одного и того же языка не всегда говорят одинаково, на их речь влияет множество лингвистических и экстралингвистических факторов.

Об особенностях речи удобно говорить с точки зрения коммуникативной ситуации, т. е. ситуации, в которой протекает речевая коммуникация. В результате взаимодействия и изменения различных факторов коммуникативной ситуации возникают различные идиомы. В зависимости от характера этих факторов можно выделить идиомы, которые определяются по социальным характеристикам говорящего, и стилистические разновидности языка [Ерофеева, 2005, с. 37].

Важными понятиями, которые здесь стоит упомянуть, являются стиль произношения и тип произнесения [Бондарко и др., 1974]. Термин «стиль произношения» восходит к Л. В. Щербе, однако сейчас принято разграничивать эти два понятия. Под стилем произношения понимается общая фонетическая характеристика высказывания, тогда как под типом произнесения — детальная фонетическая характеристика элементов речи. Полный тип произнесения подразумевает возможность однозначной фонемной интерпретации отрезка речи; соответственно, неполный — невозможность такой интерпретации. Л. В. Бондарко указывает, что в речевой цепи сегменты полного и неполного типа произнесения обычно чередуются. В спонтанной речи процент сегментов неполного типа значителен.

Нельзя не упомянуть и диалектные и региональные различия, которые также проявляются на фонетическом уровне. Множество исследований посвящены фонетическим особенностям русских диалектов [Князев, 2008] и региолектов [Ерофеева, 2005].

Наконец, вариативность может быть обусловлена и анатомическими факторами, например, разницей в длине речевого тракта. Кроме того, на речь

могут влиять внешние факторы среды, например, шумная обстановка [Lombard, 1911].

2.3. Фонетические особенности русской разговорной речи

Описанию различных фонетических особенностей русской разговорной речи посвящено множество работ; здесь будет кратко обобщён ряд таких работ, результаты которых представляются перспективными с точки зрения составления произносительных словарей. Стоит также отметить, что из этих же соображений будут рассмотрены работы, касающиеся сегментных особенностей. Наконец, здесь не будут рассмотрены работы, касающиеся диалектных или региональных особенностей, так как данный вопрос выходит за рамки исследования.

В исследовании [Bondarko и др., 2003] на материале корпуса спонтанной русской речи INTAS выявлен ряд особенностей. Так, было показано, что форманты гласных в спонтанной речи проявляют гораздо большую вариативность, а также подтверждалась точка зрения, что предударные и заударные гласные составляют по своим акустическим характеристикам две различные группы. Кроме этого, часто наблюдались гласные вставки в кластерах согласных, а также выпадения безударных (особенно заударных гласных) с сохранением слоговой структуры (в этих случаях роль слогообразующего элемента принимает на себя сонант, звонкий согласный или глухой щелевой согласный).

Работа [Болотова, 2005] посвящена экспериментально-фонетическому исследованию гласных в чтении и в спонтанной речи. Было показано, что основной особенностью связного текста является значительная количественная и качественная редукция. Гласные /i/ и /u/ оказались наиболее устойчивыми по качественным характеристикам, тогда как /a/ — самым неустойчивым; также этот гласный оказался наиболее вариативным с количественной точки зрения. Показано также, что безударные гласные в спонтанной речи чаще выпадают,

чем заменяются на какой-либо иной звукотип. Некоторые особенности, по-видимому, являются дикторозависимыми характеристиками.

Исследование [Скрелин, Евдокимова, 2008] посвящено аллофонному варьированию гласных в спонтанной речи. Было показано, что часто встречается реализация на месте определённых фонем аллофонов, принадлежащих другой фонеме. Такие замены наиболее характерны для фонемы /i/. Количество реализаций аллофонов фонем /e/, /u/ и /a/ на месте аллофонов других фонем доходило до 15 % исследуемого материала.

Работа [Васильева, Тананайко, 2009] посвящена морфонологическим факторам, влияющим на редукцию безударных гласных. Показано, что аллофоны фонемы /a/ в первообразных предлогах во втором предударном слоге имеют тенденцию реализоваться как аллофоны этой же фонемы, обычно встречающиеся в первом предударном слоге.

Исследование [Evgrafova, 2009] посвящено гласным вставкам в консонантных кластерах. На материале чтения вслух изолированных слов показано, в каких контекстах встречаются гласные вставки, а также проанализированы акустические характеристики такого гласного звука (F_1 от 400 до 500 Гц, F_2 от 1300 до 1500 Гц).

В исследовании [Васильева, Тананайко, 2010] рассматриваются заударные звуковые последовательности. В таких позициях был обнаружен ряд изменений. Так, спирантизации были подвержены согласные в комплексе /tʃisk/ в словах типа «практически» (/tʃ/ заменялся на /ʃʲː/ или /ʒ/) и в комплексе /tʲilʲ/ в словах типа «подготовительное» (/tʲ/ заменялся на /sʲ/, /zʲ/ или [θ]). Фонема /v/ способна реализоваться как аппроксиманты [ʊ], [w] или [ɥ]. Носовые согласные /m/ и /n/ могут реализовываться как неопределённый носовой гласный. Также был выявлен ряд случаев, в которых выпадают звуки или группы звуков. Так, в указанных выше комплексах выпадает гласный /i/.

Исследование [Апушкина, 2011] посвящено особенностям качества гласных в спонтанной речи. Показано, что ударные гласные характеризуются специфическим набором акустических признаков, в числе которых не представлено отсутствие качественной редукции.

В работах [Ронжин, Евграфова, 2011; Ронжин, Евграфова, Кипяткова, 2011] обозреваются исследования, рассматривающие особенности спонтанной речи и факторы, которые влияют на её вариативность. В частности, упомянута монография [Андросова и др., 2006], показывающая, что наблюдается корреляция между актуальным членением предложения и чередованием сегментов полного и неполного типа произнесения.

В работе [Горлова, Слепокурова, 2012] изучаются особенности редукции звуков в предударных позициях. Показывается, что в более 90 % случаев потерь в сегментном составе не происходит. Однако проводится анализ найденных явлений: так, показано, что зияния могут подвергаться стяжению или элизии, причём наиболее яркие стяжения образуются при начальном гласном /i/.

Исследование [Венцов и др., 2013] представляет корпус русской спонтанной речи с профессиональной фонетической транскрипцией и созданный на его основе частотный словарь спонтанных словоформ.

Работа [Nigmatulina, 2013] рассматривает стяжения звуков в спонтанной речи. Показано, что стяжение гласных равновероятно на границах слов и внутри слова, в то время как согласные больше подвержены стяжению внутри слов. Чаще всего гласный, получившийся в результате стяжения, был одним из исходных гласных, но были и случаи появления качественно иного звука. Место ударения также влияет на наличие или отсутствие стяжения: ударные звуки реже, чем безударные, подвергаются стяжению внутри слова и несколько чаще — на границах. Согласные одного места образования подвергаются ассимиляции, причём результирующий звук, хотя и обладает большей длительностью, не воспринимается как удвоенный.

В работе [Кочаров, Кочеткова, 2020] исследуется ассимиляция русских безударных гласных по огубленности. Показано, что аллофон фонемы /а/ второй степени редукции [ə] подвергается огублению в позициях перед огубленным гласным /u/, например, в слове «голубой», однако этого не происходит с аллофоном первой степени редукции [ʌ]: в слове «**а**ккуратный» первый звук [ʌ] огубляться не будет, однако огубление может происходить в сочетании «к **а**ккуратному», где фонема /а/, не будучи в абсолютном начале, реализуется аллофоном второй степени редукции.

2.4. Особенности реализации окончаний прилагательных

Одной из произносительных особенностей, на которую стоит обратить особое внимание, является реализация окончаний прилагательных. Ряд падежных окончаний прилагательных в качестве своей морфемной оболочки имеет два гласных, разделённых фонемой /j/. Однако эта фонема имеет тенденцию выпадать в интервокальной позиции перед безударными гласными, из-за чего образуется сочетание «гласный+гласный», т. е. зияние. Реализация таких сочетаний характеризуется сильными взаимным влиянием гласных, что служит причиной образования хорошо слышимого переходного участка и значительных формантных изменений гласных [Бондарко, 1977, с. 105]. Кроме того, часто эти окончания находятся в заударной позиции, что также заметно влияет на фонетическое качество звуков [Бондарко, 1977, с. 155]. Поскольку правильное определение окончания обеспечивает правильное определение формы слова и, как следствие, правильную грамматическую организацию гипотезы системы распознавания речи, целесообразны попытки улучшить распознаваемость таких окончаний.

2.6. Выводы по главе 2

Итак, в главе 2 были рассмотрены фонетические особенности русского языка, которые могут быть отражены в системе АРР — как общеязыковые, так и подверженные внутридикторской и междикторской вариативности. В

практической части исследования будет рассмотрена возможность отражения ряда особенностей в произносительном словаре.

Фонетические характеристики, выявленные на материале корпуса INTAS, отражены в его аннотации, что позволяет их выделение и использование, в том числе как материала для машинного обучения. Перспективным направлением исследований представляется также отражение в словарях результатов фонетических изменениях, происходящих в сочетаниях «гласный+гласный» и «гласный+/j/+гласный» в окончаниях форм прилагательных (в том числе стяжений), чему будет посвящён отдельный эксперимент.

Таким образом, материал, представленный в данной главе, представляет собой теоретическую базу для проводимых исследований.

Глава 3. Автоматическое создание произносительного словаря

3.1. Материал исследования

Материалом для исследования послужило несколько корпусов устной русской речи, разработанных и записанных на кафедре фонетики и методики преподавания иностранных языков СПбГУ.

Первым из них стал корпус INTAS, представленный в 2003 году [Bondarko и др., 2003]. Корпус включает в себя речь десяти дикторов, а именно: запись спонтанного монолога; запись чтения вслух текста, лексически совпадающего со спонтанным монологом; запись чтения вслух двух фонетически представительных текстов; а также запись чтения отдельных слов. Подкорпус спонтанной речи включает в себя 44 минуты речи, 6431 словоупотребление, 2026 словоформ, 1430 лексем, чтения — 36 минут речи, 5421 словоупотребление, 1814 словоформы, 1297 лексем. Корпус обладает многоуровневой разметкой. В уровни среди прочих входят уровень идеальной фонемной транскрипции, уровень реальной фонетической транскрипции, уровень обозначения ударности слога и уровень орфографической расшифровки. Транскрипции и разметка были составлены профессиональными фонетистами. Пример аннотации корпуса приведён на рис. 4.

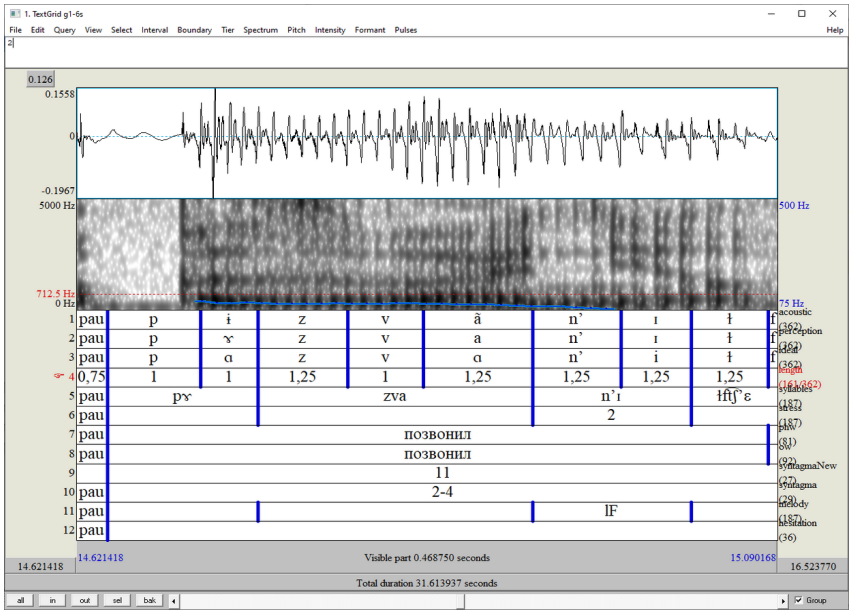

Рис. 4 — Фрагмент аннотации корпуса INTAS

Ещё одним корпусом, послужившим материалом для исследования, стал корпус CORPRES (корпус профессионального чтения) [Skrelin и др., 2010]. В данный корпус входят записи чтения вслух 8 профессиональными дикторами нескольких художественных текстов. Корпус насчитывает более 60 часов речи, 211 389 словоупотреблений, 16 226 словоформ, 9815 лексем. Корпус также обладает многоуровневой разметкой, которая включает уровень орфоэпической фонемной транскрипции, уровень реальной фонемной транскрипции и уровень орфографической расшифровки. Орфоэпическая транскрипция была создана автоматически с ручной коррекцией, реальная транскрипция и расстановка границы звуков были проведены экспертами-фонетистами. Пример аннотации корпуса приведён на рис. 5.

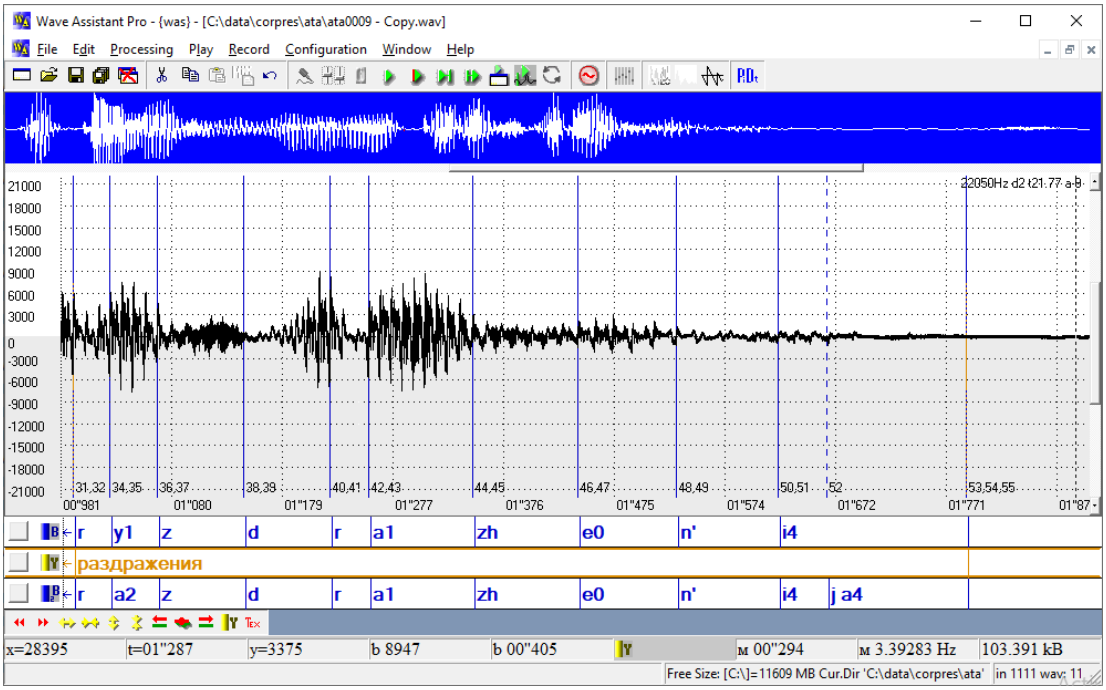

Рис. 5 — Фрагмент аннотации корпуса CORPRES

Наконец, самым новым корпусом в материале исследования стал корпус диалоговой речи SibLing [Kachkovskaia и др., 2020]. Корпус содержит в себе записи 90 игровых диалогов между 100 дикторами, различными по степени знакомства, возрасту, полу и положению в обществе. Общая длительность корпуса составляет 65 часов диалогической речи. Разметка корпуса включает в себя орфографическую расшифровку и фонемную транскрипцию, созданную автоматически по правилам [Евдокимова, Скрелин, Чукаева, 2017]. В данном корпусе используется его подкорпус из 72 диалогов. Данный подкорпус насчитывает 30 часов диалогической речи, 114 865 словоупотреблений, 6754 словоформы, 3389 лексем.

3.2. Инструменты для автоматического создания транскрипций

Ряд экспериментов в рамках данного исследования был посвящён вопросам создания транскрипций при помощи инструментов машинного обучения. Было решено использовать и сравнить два инструмента: конечные автоматы (а именно взвешенные конечные преобразователи, finite-state

transducers, FST) и искусственные нейронные сети (а именно двунаправленные сети с долгой краткосрочной памятью, Bidirectional Long Short-Term Memory Neural Networks, BiLSTM). Оба эти инструмента традиционно используются для задач моделирования временных последовательностей, в том числе и автоматического создания транскрипций (G2P).

3.2.1. Взвешенные конечные преобразователи

В качестве инструментария для автоматического создания транскрипций при помощи FST был взят пакет Phonetisaurus [Novak, Minematsu, Hirose, 2016], основанный на библиотеке OpenFST [Allauzen и др., 2007]. Данный пакет успешно применялся для задач G2P на различных языках (английский [Novak, Minematsu, Hirose, 2016], арабский [El-Hadi, Mhania, 2017], румынский [Domokos, Szakács, 2017]).

Подход, использованный авторами пакета, заключается в следующем [Novak, Minematsu, Hirose, 2016]: на первом этапе проводится выравнивание обучающего материала, т. е. определение того, какой букве (или буквосочетанию) какой звук соответствует. Алгоритм выравнивания также основан на конечных преобразователях; основная идея заключается в том, чтобы из всех возможных выравниваний на корпусе найти наиболее вероятные. Пример такого выравнивание показан в таблице 3.

Таблица 3 — Пример выравнивания транскрипций слов «мел» и «мель» (символом # обозначена «нулевая» фонема)

м	е	л		м	е	л	ь
m^j	é	l		m^j	é	l^j	#

На следующем этапе материал используется для создания стандартной n-граммной модели, которая затем преобразуется во взвешенный конечный преобразователь. Пересечение представления слова в виде конечного

преобразователя и n-граммной модели даст граф, наиболее вероятный путь по которому и будет представлять искомую транскрипцию для слова. Такой алгоритм даёт возможность для каждого слова давать не одно произношение, а несколько наиболее вероятных.

3.2.2. Двунаправленные сети с долгой краткосрочной памятью

Долгая краткосрочная память (long short-term memory) — особая архитектура рекуррентных нейронных сетей, которая может обрабатывать целые последовательности элементов данных [Hochreiter, Schmidhuber, 1997]. В отличие от обычных рекуррентных ИНС, LSTM лучше запоминают данные с большими промежутками между ними. Двунаправленный вариант подразумевает запоминание не только левого контекста, но и правого. Это позволяет моделировать как регрессивные, так и прогрессивные коартикуляционные процессы и может послужить преимуществом в сравнении с WFST.

Для экспериментов с искусственными нейронными сетями была взята реализация из модуля Tensorflow [Dillon и др., 2017] для языка программирования Python. Данный модуль обладает широкими возможностями для обучения и применения нейронных сетей различной архитектуры. В данном исследовании будет рассмотрен вариант одной из архитектур, работоспособность которой для данных задач была показана специалистами из компании Google [Rao и др., 2015]. Такая нейронная сеть включает себя два слоя BiLSTM с 256 единицами памяти и сигмоидной функцией активации и выходной полносвязный слой с функцией активации softmax.

На вход нейросети подаются последовательности символов, закодированных с помощью унитарного прямого кодирования (one-hot encoding, OHE) [Harris, Harris, 2010, с. 129]: это значит, что каждому символу сопоставляется вектор размерности N (где N — размер входного алфавита), в котором все элементы равны 0, кроме одного — с индексом, соответствующим

порядковому номеру символа в алфавите. Проиллюстрируем концепцию с помощью упрощённого примера: если алфавит состоит из символов «а», «б» и «в», то они будут кодироваться векторами размерности 3 — [1, 0, 0], [0, 1, 0] и [0, 0, 1] соответственно.

Таким образом, входная размерность нейросети составит L x N, где L — длина входящей последовательности, N — размер входного алфавита. Выходной же слой нейросети будет иметь размерность M, где M — размер выходного алфавита. Выходами нейросети будут символы транскрипции, также закодированные с помощью OHE.

Нейросеть компилируется при помощи оптимизатора Adam, функция loss (которая описывает разницу между предсказанием системы и реальным ответом: чем ниже loss, тем выше качество системы), использующаяся в процессе обучения — категориальная кросс-энтропия, обыкновенно использующаяся для категориальных данных, какими и являются текстовые символы.

В процессе обучения данные разделяются на обучающее множество (90 % данных) и валидационное (10 % данных). Обучение проходит итеративно: на каждой итерации (эпохе) обучающее множество перемешивается и пакетами (batch) подаётся на вход нейросети. В конце каждой эпохи для контроля обучения функция loss вычисляется на валидационном множестве: это делается для того, чтобы избежать переобучения, когда модель повышает точность предсказания на обучающем множестве, идеально его описывая, за счёт понижения точности на новых для неё данных. В каждом эксперименте обучение останавливалось, когда функция loss переставала падать (т. е. качество предсказания на валидационной выборке прекращало расти).

Кроме функции loss, для контроля использовалась функция категориальной точности, которая отражала процент «угаданных» системой ответов. Чем выше эта метрика, тем выше качество системы.

3.3. Инструменты для автоматического распознавания речи

Все созданные в данной работе системы распознавания речи были построены на основе инструментария для распознавания речи Kaldi, созданного и разрабатываемого Дэниелом Поуви (Daniel Povey) в университете Джона Хопкинса, Балтимор, Мэриленд, США, и в Брненском техническом университете, Брно, Чехия [Povey и др., 2011]. Этот инструментарий предлагает набор алгоритмов и рецептов для распознавания речи и широко используется как в научных исследованиях, так и в коммерческих целях. Представленные в Kaldi алгоритмы используют скрытые Марковские модели (HMM) и глубокие нейронные сети (DNN). В связи с ограниченным речевым материалом были использованы только алгоритмы, основанные на скрытых Марковских моделях.

Код Kaldi написан на языке программирования C++, а внешние скрипты, регулирующие работу алгоритмов, написаны на языках Perl и Bash.

Для создания языковых моделей в Kaldi используется инструментарий SRILM (SRI Language Modeling Toolkit), разработанный в лаборатории речевых технологий института SRI International (SRI Speech Technology and Research Laboratory) [Stolcke, 2002].

3.4. Использованное оборудование

Для экспериментов с искусственными нейронными сетями использовался персональный компьютер со следующими характеристиками:

Операционная система: Windows 7 x86-64

Процессор: Intel Core i7-5930K CPU @ 3.50GHz (тактовая частота 3.50 ГГц)

Графический процессор: NVIDIA GeForce RTX 2080 Ti (для проведения вычислений на графическом процессоре использовались архитектура CUDA и библиотеки для глубокого обучения cuDNN, что позволило значительно ускорить процесс обучения нейронных сетей).

ОЗУ: 16.00 Гб.

Для экспериментов со взвешенными конечными преобразователями и с распознаванием речи использовался персональный компьютер со следующими характеристиками:

Операционная система: Ubuntu 20.04.1 64-бит

Процессор: Intel Core i5-6500 CPU @ 3.20GHz (тактовая частота 3.19 ГГц)

ОЗУ: 8.00 Гб.

3.5. Автоматическое создание транскрипций

3.5.1. Автоматическое создание идеальных транскрипций

Первая часть экспериментов была посвящена созданию фонемных транскрипций. Испытывались различные комбинации обучающего материала и использованного инструмента.

Каждый обучающий набор состоит из пар: первым элементом такой пары является последовательность «входных» символов (например, орфографических), вторым — последовательность «выходных» символов (например, символов орфоэпической транскрипции). Пример одной строчки обучающего корпуса приведён в таблице 4.

Таблица 4 — Пример элемента обучающего материала

X	и та́м я начина́ю маршру́т
Y	i tám ja nat͡ʃináju marʃrút

Из корпуса CORPRES были выделены все доступные словоупотребления, последовательности слов от паузы до паузы и их орфоэпические транскрипции составили обучающий материал, обозначенный как «CORPRES — идеальная транскрипция». В корпусе SibLing обучающий материал был организован аналогичным образом, однако фрагменты из речи одного диктора, взятой из двух диалогов, общим объёмом 3159 словоупотреблений были изъяты из обучающего материала и использовались как тестовый. Таким образом, было

получено два обучающих набора, каждый из которых был протестирован на обеих рассматриваемых системах. Метрикой для сравнения гипотезы системы и эталона был выбран процент неправильно предсказанных звуков — Phoneme Error Rate. Результаты эксперимента приведены в таблице 5.

Таблица 5 — Эксперименты по автоматическому созданию идеальных транскрипций

Обучающий материал	**Инструмент**	**Phoneme Error Rate**
CORPRES — идеальная транскрипция	Phonetisaurus	8.75 %
SibLing — автоматическая транскрипция	Phonetisaurus	1.74 %
CORPRES — идеальная транскрипция	BiLSTM	6.44 %
SibLing — автоматическая транскрипция	BiLSTM	2.75 %

Видно, что FST и нейронная сеть показывают разные результаты для разного обучающего материала. Вероятно, конечные автоматы лучше усваивают наборы правил, на которых основан транскриптор (за счёт этого Phonetisaurus показывает лучший результат на материале автоматической транскрипции, чем нейронная сеть). Однако нейронная сеть совершает меньше ошибок в случае обучения на материале CORPRES, что даёт основания для предположения о большей сложности данного материала: «обычные» конечные автоматы хуже улавливают правила транскрипции.

Ручное сравнение результатов экспериментов позволило сравнить методы транскрипции, использованные в корпусах SibLing и CORPRES и которым, соответственно, обучились системы. В таблице 6 показана краткая сводка такого анализа.

Таблица 6 — Сравнение систем транскрипции в корпусах

Орфография	**CORPRES**	**SibLing**
двойной согласный *военного*	/vajé**n**ava/	/vajé**nn**ava/
выпадение фонемы /j/ *мавзолеем*	/mavzalʲé**j**im/	/mavzalʲéim/
ассимиляция по мягкости *поднялась*	/pa**dn**ʲilás'/	/pa**d**ʲ**n**ʲilás'/
ассимиляция по звонкости *есть два*	/ji**z**ʲ**d**ʲ dvá/	/ji**s**ʲ**d**ʲ dvá/

В первых трёх из перечисленных различий оба варианта имеют право на существование, отражая разные произносительные варианты и уровни абстракции. Однако в последнем возможен только один вариант, а вариант, предложенный автоматическим транскриптором, приходится считать ошибочным. В этом несложно заметить недостаток систем, основанных на правилах: при большом их количестве человеческий фактор может позволить упустить какое-то их количество, порождая ошибки.

Перечисленные различия системны для обучающего материала, что объясняет, почему обучение на материале SibLing последовательно даёт лучшие результаты: обе системы усваивают методы, использованные в обучающем материале, и для системы на основе CORPRES эти методы не совпадают с «эталонными».

Анализ ошибок, совершённых нейронной сетью, показал, что в случае обучения на CORPRES система в некоторых случаях случайным образом удваивает или утраивает звуки, что может отразиться и на соседних звуках, например: «чуть дальше» — /t͡ʃut' á**ʃʃʃ**/, «заозерье» — /zaa**éé**rʲji/, чего не наблюдается на корпусе SibLing; вероятно, это можно объяснить наличием небольшого количества шума в обучающих данных.

Взвешенные конечные автоматы таких ошибок не совершали, однако было замечено, что в случае обучения на CORPRES система могла предсказывать выпадение фонемы /j/ в случаях, где это маловероятно: маяк — /maák/. Такие ошибки также можно объяснить шумом в обучающих данных.

Отдельно стоит заметить, что поскольку в орфографической аннотации корпуса CORPRES не использовалась буква «ё», но она использовалась в корпусе SibLing, системы, обученные на CORPRES, требуют предобработку данных в виде замены «ё» на «е». Однако это не мешает системе на FST правильно определять случаи, где букве «ё» соответствует фонема /o/ (т. е. должна была быть буква «ё»). Слово «черёмуха» ни разу не встретилось в корпусе CORPRES, тем не менее, Phonetisaurus предлагает транскрипцию /t͡ʃirʲómuxa/ для последовательности букв «чере́муха» (тогда как нейронная сеть — /t͡ʃirʲémuxa/).

3.5.2. Автоматическое создание реальных транскрипций

На данном этапе исследования было рассмотрено две стратегии: одна подразумевала переход от орфографии к реальной транскрипции, другая — переход от идеальной транскрипции к реальной. Таким образом, по методике, аналогичной предыдущему эксперименту, были сформированы следующие обучающие наборы:

1. Аналогичный набору «CORPRES — идеальная транскрипция», но с использованием реальной фонемной транскрипции из аннотации корпуса.

2. Набор, в котором «входными» данными была орфоэпическая фонемная транскрипция корпуса CORPRES, «выходными» — реальная фонемная.

3. Набор, в котором «входными» данными была орфоэпическая фонемная транскрипция корпуса INTAS, «выходными» — реальная фонетическая.

Далее, как и в предыдущем эксперименте, для каждой конфигурации была обучена BiLSTM-модель по тем же алгоритмам обучения.

Исследованные комбинации приведены в таблице 7. Для каждого случая указана точность предсказания на валидационном множестве на последней итерации обучения.

Таблица 7 — Эксперименты по автоматическому созданию реальных транскрипций

Обучающий материал — ввод	**Обучающий материал — вывод**	**Инструмент**	**Точность**
CORPRES — орфография	CORPRES — реальная транскрипция	BiLSTM	87.01 %
CORPRES — идеальная транскрипция	CORPRES — реальная транскрипция	BiLSTM	84.67 %
INTAS — идеальная транскрипция	INTAS — реальная транскрипция	BiLSTM	95.15 %

Наиболее успешно показала себя система на основе INTAS, тогда так точность систем на основе корпуса CORPRES несколько ниже.

Поскольку для данного типа транскрипции невозможно сравнение с автоматическим транскриптором (т. к. он не отражает интересующих нас данных), было принято решение проверить качество созданных таким образом транскрипций путём их практического применения в качестве произносительных словарей для распознавания тестового фрагмента корпуса

SibLing — т. е. необходимо выяснить, какая из систем даст наивысшее качество автоматического распознавания речи. Эксперимент был построен следующим образом:

1. На основе результатов обработки тестового фрагмента каждым из транскрипторов были созданы произносительные словари. Использовались орфографическая запись фрагмента, его изначальная фонемная транскрипция, созданная по правилам, и его фонемная транскрипция, созданная на основе транскриптора, обученного на уровне орфоэпической транскрипции из корпуса CORPRES.

2. Тестовый материал был разделён на две части, одна из которых послужила обучающим материалом для системы APP, а другая — тестовым.

3 На всём материале была обучена языковая модель; использования такой модели позволяет свести к минимуму все ошибки, которые могут быть вызваны её несовершенством [Холявин, 2019].

4. На обучающем материале для каждого словаря была обучена акустическая модель HMM-GMM.

5. С помощью каждой модели было проведено распознавание тестового фрагмента и вычислен параметр WER.

Результаты распознавания приведены в таблице 8.

Таблица 8 — Эксперименты по автоматическому распознаванию речи

Произносительный словарь	**WER**
SibLing — изначальная автоматическая транскрипция	14.00 %
CORPRES — идеальная транскрипция	14.14 %
CORPRES — реальная транскрипция (на основе орфографии)	13.40 %

CORPRES — реальная транскрипция (на основе изначальной фонемной транскрипции)	13.30 %
CORPRES — реальная транскрипция (на основе транскрипции, созданной с помощью МО)	14.08 %
INTAS — фонетическая транскрипция (на основе изначальной фонемной транскрипции)	16.14 %

Все рассмотренные системы показывают близкие результаты, что говорит об их применимости к данной задаче. Несмотря на то, что система на основе INTAS показывала лучшие результаты в процессе обучения, она оказалась несколько менее пригодна для задач автоматического распознавания речи. Вероятно, модель на основе большего количества обучающего материала показала бы лучшие результаты.

Наилучшие же результаты были показаны на материале словарей, обученных на материале реальной фонемной транскрипции из корпуса CORPRES. Можно предположить, что использование такой реальной транскрипции является компромиссными вариантом между орфоэпической транскрипцией (не отражающей реальных фонетических процессов) и точной фонетической (вероятно, слишком подробной и вносящей в систему фактор неопределённости). Такую транскрипцию можно генерировать как напрямую из орфографии, так и на основе орфоэпической фонемной транскрипции, созданной по правилам. Генерация же транскрипции на основе фонемной, также полученной с помощью методов машинного обучения, даёт несколько худшие результаты, что, вероятно, объясняется накоплением ошибок, совершённых обеими системами.

Орфоэпическая транскрипция на материале корпуса CORPRES показала очень близкие результаты к изначальной автоматической фонемной

транскрипции; вероятно, это говорит о том, что несмотря на освещённые выше различия между системами, они обе одинаково адекватны и в равной степени пригодны для задач АРР.

Анализ ошибок, совершённых системой в каждом случае, не позволил выявить каких-либо системных различий. Во всех случаях ошибки в основном были связаны с неправильным распознаванием формы слова («администрации» вместо «администрацию», «огибаем» вместо «огибаю»), вставками коротких слов («и», хезитация), пропуском коротких слов («как тебя получилось» вместо «как у тебя получилось»). Можно предположить, что первый тип ошибок может быть исправлен при введении в систему альтернативных транскрипций для окончаний, в которых и заключается ошибка. Второй и третий тип ошибок, возможно, будет исправлен при улучшении языковой модели таким образом, чтобы она не позволяла ошибочных с точки зрения грамматики словосочетаний.

3.6. Анализ необходимого объёма данных для обучения системы G2P

Отдельный эксперимент был посвящён исследованию зависимости между объёмом обучающих данных и качеством транскрипции, созданной с помощью взвешенных конечных автоматов (как и ранее, с использованием пакета Phonetisaurus). В качестве обучающего материала были взяты все транскрипции из корпуса CORPRES, в качестве тестового — фрагмент художественного текста. За эталон была взята транскрипция фрагмента, созданная на основе правил. В ходе эксперимента были созданы модели, основанные на материале всего корпуса, а также на фрагментах меньшего объёма, начиная от 10 словоупотреблений. Затем с помощью каждой из обученных моделей были составлены транскрипции тестового фрагмента, которые были сравнены с эталоном при помощи параметра Phoneme Error Rate (PER).

Сравнение показало, что модель, обученная на полном корпусе, даёт PER, равный 4,16 %. Модель, обученная всего на 10 словоупотреблениях, дала ошибку в 84,43 %. Обучение на 50 словоупотреблениях понизило её до 24,01 %,

на 100 словоупотреблениях — до 19,81 %, на 500 словоупотреблениях — 11,50 %, на 1000 — 9,68 %, на 15 000 — 7,56 %, и далее наблюдалось плавное снижение. Таким образом, зависимость качества транскрипции от количества обучающего материала носит нелинейный характер. График зависимости приведён на рис. 6.

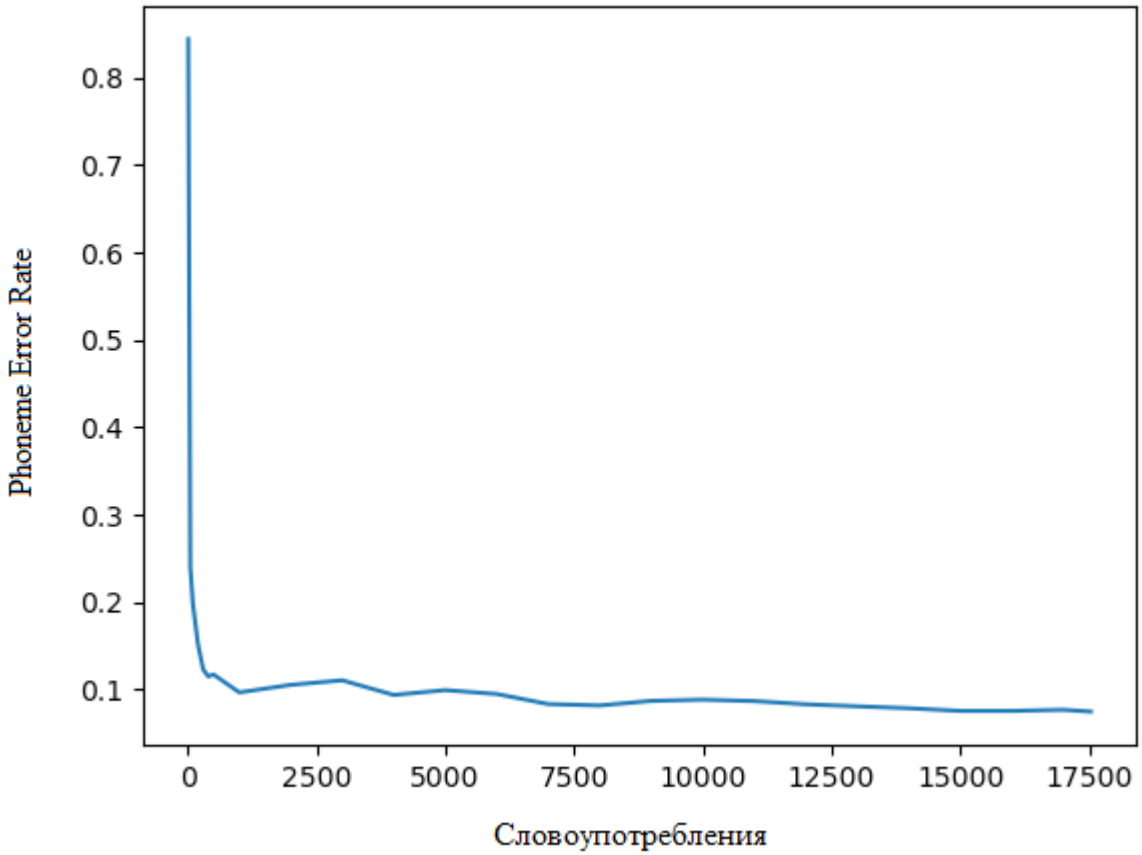

Рис. 6 — Зависимость качества системы G2P от количества обучающего материала

3.7. Использование произносительных вариантов для морфем

Результаты анализа ошибок системы, приведённые в разделе 3.5.2, позволили предположить, что введение в систему альтернативных вариантов

транскрипции для окончаний словоформ может повысить качество распознавания. Поэтому в данном разделе рассматривается возможность введения таких вариантов. Как уже было указано выше, фонетическая реализация окончаний прилагательных в русском языке значительно отличается от идеальной транскрипции, поэтому они были выбраны в качестве материала эксперимента. Эксперимент был поставлен следующим образом:

1. На материале подкорпуса «а» корпуса CORPRES (где проставлены границы между звуками; данный подкорпус содержит 79 102 словоупотребления) с помощью пакета pymorhpy2 для языка программирования Python [Korobov, 2015] были выделены прилагательные (морфологический тег «ADJF», т. е. полная форма прилагательного), из которых были выбраны формы с окончаниями «ая», «ее», «ие», «ое», «ую», «ые», «юю», «яя» (всего 1369 словоупотреблений). Эти окончания не содержат в себе других согласных (кроме /j/, который склонен к выпадению в интервокальной позиции), а значит, в них будут встречаться зияния.

2. Остальной материал, не включающий таких прилагательных, был взят в качестве обучающего. На его основе была создана акустическая модель HMM-GMM.

3. На том же материале была обучена особая 5-граммная языковая модель, где каждое слово состояло только из одного звука. Таким образом, модель отражала фонотактические особенности языка, и в сочетании с акустической моделью составляла систему пофонемного распознавания речи.

4. С помощью такой системы тестовый материал (т. е. формы прилагательных) был распознан и были выделены гипотезы системы относительно реального произношения окончаний прилагательных.

5. Для проверки целесообразности использования таких транскрипций была создана система поморфемного распознавания с несколькими вариантами произносительного словаря: в одном транскрипции окончаний были

идеальными, в другом использовались гипотезы системы пофонемного распознавания, другие представляли собой промежуточные варианты.

В таблице представлены результаты распознавания для каждого произносительного словаря.

Таблица 9 — Разные транскрипции окончаний

Словарь	**Morpheme Error Rate**
1. Идеальная транскрипция	11.32 %
2. Гипотезы + идеальные + варианты из корпуса	5.84 %
3. Гипотезы	8.69 %
4. Гипотезы (с ограничением по встречаемости)	9.06 %
5. Вар. 2, но без транскрипций с /j/	10.30 %

Наилучшие результаты показал словарь с наибольшим количеством произносительных вариантов для каждого окончания. Он включал в себя как идеальные, так и альтернативные произношения для каждой морфемы. Все рассмотренные варианты словарей приведены в приложении.

Использование исключительно орфоэпических произношений для каждого окончания даёт наихудший из рассмотренных результатов распознавания; этот словарь подразумевает наименьшее разнообразие произносительных вариантов. Словарь же с наибольшим разнообразием, куда входили как орфоэпические, так и упрощённые произношения, даёт, напротив, наилучший результат. Вариант словаря 3 даёт несколько лучший результат, чем вариант 4, входящий в него, что также подтверждает гипотезу о важности разнообразия транскрипций. Наконец, низкие результаты показывает вариант,

где во всех произношениях отражено выпадение /j/ в интервокальном положении перед безударным гласным, кроме /u/. Вероятно, это обусловлено тем, что несмотря на перцептивное выпадение фонемы, её фонетическое влияние на контекст остаётся достаточно сильным, чтобы быть значимым для системы автоматического распознавания речи. Таким образом, можно сделать вывод, что наилучшими словарями будут те, в которых отражены все возможные произношения, в том числе и полные орфоэпические.

3.8. Выводы по главе 3

В данной главе приведены описания и результаты экспериментов, направленных на поиск методов оптимизации произносительных словарей.

Эксперименты с взвешенными конечными преобразователями и искусственными нейронными сетями показали, что для создания произносительных словарей целесообразно применение методов машинного обучения; как конечные автоматы, так и нейронные сети успешно обучаются транскрипционным правилам, по которым был составлен их обучающий материал. Нейронные сети, вероятно, более приспособлены для работы с более сложными правилами. Доступный обучающий материал (сотни тысяч словоформ) вполне достаточен для обучения систем создания орфоэпических фонемных транскрипций и сильно превышает минимальный необходимый (тысячи словоформ). Однако вероятно, что минимальный объём для создания точных фонетических транскрипций существенно выше, и существующих на данный момент данных (тысячи словоформ) недостаточно для создания системы, которая бы внесла существенный вклад в качество APP.

Эксперименты по применению систем G2P для создания словарей для автоматического распознавания речи показали, что их использование возможно для генерации словарей, основанных на разных типах транскрипции. Использование автоматически сгенерированной реальной фонемной транскрипции, как на основе орфографической записи, так и на основе

орфоэпической фонемной транскрипции, вносит улучшение в качество распознавания по сравнению с использованием исключительно последней.

Тем не менее, не удалось добиться улучшения качества распознавания при использовании словарей, основанных на точной фонетической транскрипции. Такие словари показали наихудшие результаты из всех протестированных.

Вероятно, лучшие результаты покажут системы, обученные на большем количестве входных данных: реальные фонетические транскрипции, точно описывающие произносимое, не являются частью большинства доступных корпусов по причине сложности и времязатратности. Только один корпус из использованных трёх обладал такой транскрипцией, и его объём существенно меньше по сравнению с двумя другими. Таким образом, приходится признать, что основной проблемой в области создания точных фонетических транскрипций является недостаток качественных обучающих данных.

Эксперименты с оптимизацией транскрипций конкретных морфем показали, что использование только идеальных транскрипций для этих морфем даёт значительное ухудшение распознавания. Однако использование только редуцированных транскрипций также даёт результаты значительно хуже тех, которые достигаются с использованием словаря с наибольшим разнообразием транскрипций. Вероятно, аналогичных результатов можно добиться и в других схожих контекстах, таких как окончания причастий, местоимений, зияния и сочетания вида «гласный+/j/+гласный» в других контекстах, например, в корнях слов. Также стоит отметить, что в ходе экспериментов была проиллюстрирована возможность создания системы поморфемного распознавания русской речи.

Практическим результатом проведённых экспериментов стали обученные модели FST и BiLSTM, способные генерировать транскрипцию по тексту, что может найти применение и в других исследованиях, не связанных с автоматическим распознаванием речи.

Заключение

В данной работе было проведено исследование методов адаптации произносительного словаря для автоматического распознавания разных типов речи. В ходе исследования были сделаны следующие выводы:

1. Для автоматического создания транскрипций можно применять методы машинного обучения. Обучающим материалом в данном случае могут стать как аннотации размеченных вручную фонетических корпусов, так и результаты работы транскрипторов, основанных на правилах; в каждом случае результирующая система будет отражать транскрипционные конвенции, использованные при создании обучающего материала. Для качественной работы такой системы необходим материал в объёме тысяч словоупотреблений.

2. Возможно также применять методы машинного обучения для генерации реальных транскрипций на основе идеальных или орфографии. Так же, как и в предыдущем случае, значительную роль играет фактор объёма доступного обучающего материала. Вероятно, в данном случае минимальный объём обучающих данных несколько выше, чем в предыдущем.

3. Созданные таким образом транскрипции можно использовать для задач автоматического распознавания речи, однако от качества модели будет зависеть и качество распознавания. Наилучшие результаты показали словари, созданные на основе реальной фонемной транскрипции из корпуса CORPRES, причём генерация транскрипции могла происходить как из орфографической записи, так и из орфоэпической фонемной транскрипции. Наихудшие же результаты показали словари, основанные на реальной фонетической транскрипции из корпуса INTAS. Это возможно объяснить как недостатком обучающего материала ввиду маленького объёма корпуса INTAS, так и повышенным фактором неопределённости системы в силу большей точности транскрипции.

4. Изменение вариантов произношения отдельных морфем в словарях может влиять на качество распознавания речи; поиск таких вариантов

представляет собой перспективное направление исследований. Показано, что для наилучшего качества распознавания окончаний форм прилагательных необходимы словари, содержащие различные варианты транскрипций, как орфоэпических, так и отражающих реальные произносительные варианты.

Полученные выводы могут служить основой для дальнейших исследований по созданию различных вариантов систем создания автоматической транскрипции для русского языка, что потенциально имеет приложения не только в области автоматического распознавания речи, но также синтеза и выравнивания (т. е. автоматической расстановки границ); последняя задача является важной для проведения многих исследований акустических особенностей языков.

Список литературы

1. Андросова С. В. и др. Реализация фонетических единиц в информационной структуре высказывания // 2006.

2. Апушкина И. Е. Качество гласных и восприятие словесного ударения в спонтанной речи // Вестник Санкт-Петербургского Университета Язык И Литература. 2011. № 4.

3. Болотова О. Б. Гласные в спонтанной речи и при чтении связного текста (экспериментально-фонетическое исследование на материале русского языка) // 2005.

4. Бондарко Л. В. и др. Стили произношения и типы произнесения // Вопросы Языкознания. 1974. Т. 2. С. 64—70.

5. Бондарко Л. В. Звуковой строй современного русского языка. : Просвещение, 1977.

6. Васильева Л. А., Тананайко С. О. Морфонологические факторы и редукция безударных гласных в нормативной русской речи // Филологические Науки Вопросы Теории И Практики. 2009. № 1. С. 32—36.

7. Васильева Л. А., Тананайко С. О. Специфика реализации заударных звуковых последовательностей в слитной речи // Филологические Науки Вопросы Теории И Практики. 2010. № 2. С. 31—33.

8. Венцов А. В. и др. Корпус русских спонтанных текстов: структура и единицы // Корпусная лингвистика-2013. , 2013. С. 223—230.

9. Горлова А. А., Слепокурова Н. А. Редукция предударных компонентов словоформ в спонтанной речи // Шестой междисциплинарный семинар «Анализ разговорной русской речи» (АРЗ-2012). , 2012. С. 15—20.

10. Евдокимова В. В., Скрелин П. А., Чукаева Т. В. Автоматический адаптивный фонетический транскриптор для русского языка // Анализ разговорной русской речи (АРЗ-2017). , 2017. С. 32—39.

11. Ерофеева Е. В. Вероятностные структуры идиомов: социолингвистический аспект. Пермь: Издательство Пермского университета, 2005. 320 с.
12. Кипяткова И. С., Карпов А. А. Модуль фонематического транскрибирования для системы распознавания разговорной русской речи // 2008.
13. Кипяткова И. С., Карпов А. А. Аналитический обзор систем распознавания русской речи с большим словарем // Труды СПИИРАН. 2010. № 12. С. 7—20.
14. Кипяткова И. С., Карпов А. А. Методология оценивания работы систем автоматического распознавания речи // Известия Высших Учебных Заведений Приборостроение. 2012. Т. 55. № 11. С. 38—43.
15. Кипяткова И. С., Карпов А. А. Разновидности глубоких искусственных нейронных сетей для систем распознавания речи // Труды СПИИРАН. 2016. Т. 6. № 49. С. 80—103.
16. Князев С. Русская диалектная фонетика // М Изд-Во МГУ. 2008.
17. Кочаров Д. А., Кочеткова У. Е. Огубленность безударных гласных в русской речи // Вопросы Языкознания. 2020. № 6. С. 31—47.
18. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук. : Российская академия наук, 1965. С. 845—848.
19. Леонтьева Ал. Б., Кипяткова И. С. Моделирование нефонемных речевых элементов и создание альтернативных транскрипций для распознавания спонтанной речи // Труды Первого Междисциплинарного Семинара «Анализ Разговорной Русской Речи»АРЗ-2007—СПб ГУАП. 2007. С. 77—85.
20. Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. М.: Издательский дом ЯСК, Рукописные памятники Древней Руси, 2016. 520 с.
21. Ронжин А. Л., Евграфова К. В. Анализ вариативности спонтанной речи и способов устранения речевых сбоев // Изв Вузов Гум Науки. 2011. Т. 2. № 3. С. 227—231.

22. Ронжин А. Л., Евграфова К. В., Кипяткова И. С. Анализ проблем автоматической обработки спонтанной русской речи // Пятый междисциплинарный семинар" Анализ разговорной русской речи"(АР3-2011). , 2011. С. 48—54.
23. Скрелин П. А., Евдокимова В. В. Вариативность реализаций гласных фонем в спонтанной речи и чтении // Второй междисциплинарный семинар" Анализ разговорной русской речи"(АР3-2008). , 2008. С. 42—47.
24. Тампель И., Карпов А. Автоматическое распознавание речи // Учебное Пособие- СПб Университет ИТМО. 2016.
25. Холявин П. А. Оценка эффективности акустических моделей для систем распознавания речи на ограниченном материале // Фонетический Лицей. 2019. № 5. С. 45—52.
26. Adda-Decker M., Lamel L. Discovering speech reductions across speaking styles and languages // Rethinking Reduction. : De Gruyter Mouton, 2018. С. 101—128.
27. Allauzen C. и др. OpenFst: A general and efficient weighted finite-state transducer library // International Conference on Implementation and Application of Automata. : Springer, 2007. С. 11—23.
28. Baker J. The DRAGON system—An overview // IEEE Trans. Acoust. Speech Signal Process. 1975. Т. 23. № 1. С. 24—29.
29. Baum L. E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes // Inequalities. 1972. Т. 3. № 1. С. 1—8.
30. Bisani M., Ney H. Joint-sequence models for grapheme-to-phoneme conversion // Speech Commun. 2008. Т. 50. № 5. С. 434—451.
31. Black A., Lenzo K. Issues in Building General Letter to Sound Rules // 2003.
32. Bondarko L. V. и др. Phonetic properties of Russian spontaneous speech // Proceedings of the 15th International Congress of Phonetic Sciences. , 2003. С. 2973—2976.

33. Bruguier A. и др. Pronunciation Learning with RNN-Transducers // Interspeech 2017. : ISCA, 2017. С. 2556—2560.
34. Bruguier A., Bakhtin A., Sharma D. Dictionary Augmented Sequence-to-Sequence Neural Network for Grapheme to Phoneme prediction // Interspeech. 2018. Т. 2018.
35. Byrne W. и др. Morpheme based language models for speech recognition of Czech // International Workshop on Text, Speech and Dialogue. : Springer, 2000. С. 211—216.
36. Casali S. P., Williges B. H., Dryden R. D. Effects of Recognition Accuracy and Vocabulary Size of a Speech Recognition System on Task Performance and User Acceptance // Hum. Factors. 1990. Т. 32. № 2. С. 183—196.
37. Chan W. и др. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). : IEEE, 2016. С. 4960—4964.
38. Dahl G. E. и др. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition // IEEE Trans. Audio Speech Lang. Process. 2011. Т. 20. № 1. С. 30—42.
39. Davel M., Wet F. Verifying pronunciation dictionaries using conflict analysis. , 2010. С. 1898—1901.
40. Davis K. H., Biddulph R., Balashek S. Automatic Recognition of Spoken Digits // J. Acoust. Soc. Am. 1952. Т. 24. № 6. С. 637—642.
41. Deri A., Knight K. Grapheme-to-phoneme models for (almost) any language // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). , 2016. С. 399—408.
42. Dillon J. V. и др. Tensorflow distributions // ArXiv Prepr. ArXiv171110604. 2017.
43. Domokos J., Szakács Z. A. Web Application for Romanian Language Phonetic Transcription // MACRo 2015. 2017. Т. 2. № 1. С. 1—10.

44. El-Hadi C., Mhania G. Phonetisaurus-based letter-to-sound transcription for Standard Arabic // 2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B). , 2017. С. 1—4.
45. Evgrafova K. The phonetic characterisics of vowel epenthesis in Russian consonant clusters // 13th International Conference on Speech and Computer SPECOM 2009. , 2009. С. 419—422.
46. Fosler-Lussier J. Dynamic Pronunciation Models for Automatic Speech Recognition // 2000.
47. Gales M. J. F., Young S. The Application of Hidden Markov Models in Speech Recognition // Found. Trends Signal Process. 2007. Т. 1. С. 195—304.
48. Gevaert W., Tsenov G., Mladenov V. Neural networks used for speech recognition // J. Autom. Control. 2010. Т. 20.
49. Giwa O., Davel M. H. Bilateral G2P accuracy: Measuring the effect of variants // 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech). Bloemfontein, South Africa: IEEE, 2017. С. 208—213.
50. Harris D., Harris S. L. Digital design and computer architecture. : Morgan Kaufmann, 2010.
51. Haykin S. Neural Networks: A Comprehensive Foundation. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. Вып. 2.
52. Hochreiter S., Schmidhuber J. Long Short-term Memory // Neural Comput. 1997. Т. 9. С. 1735—80.
53. Huang X. и др. An overview of the SPHINX-II speech recognition system. : CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 1993.
54. Huang X. и др. Spoken language processing: A guide to theory, algorithm, and system development. : Prentice hall PTR Upper Saddle River, 2001.

55. International Phonetic Association. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. : Cambridge University Press, 1999.
56. Jain A. и др. Finnish ASR with Deep Transformer Models // Conference of the International Speech Communication Association (INTERSPEECH). , 2020.
57. Jelinek F. Continuous speech recognition by statistical methods // Proc. IEEE. 1976. Т. 64. № 4. С. 532—556.
58. Jurafsky D., Martin J. H. Speech and language processing. : Pearson London, 2014.
59. Kachkovskaia T. и др. SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment // Proceeding of LREC (in press). , 2020.
60. Kaplan R. M., Kay M. Regular models of phonological rule systems // Comput. Linguist. 1994. Т. 20. № 3. С. 331—378.
61. Kessens J. M. Making a difference: On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition. : [Sl: sn], 2002.
62. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // International Conference on Analysis of Images, Social Networks and Texts. : Springer, 2015. С. 320—332.
63. Koval S., Smirnova N., Khitrov M. Modelling pronunciation variability for ASR tasks // ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology. , 2002.
64. Lama P., Namburu M. Speech recognition with dynamic time warping using MATLAB // Proj. Rep. CS 525 Spring 2010. 2010.
65. Lee K.-F. и др. Speech recognition using hidden Markov models: A CMU perspective // Speech Commun. 1990. Т. 9. № 5. С. 497—508.
66. Lombard E. Le signe de l’elevation de la voix // Ann Mal Oreille Larynx. 1911. С. 101—119.

67. Lowerre B., Reddy R. The Harpy Speech Recognition System: performance with large vocabularies // J. Acoust. Soc. Am. 1976. Т. 60. № S1. С. S10—S11.
68. Lukeš D. и др. Pronunciation Variants and ASR of Colloquial Speech: A Case Study on Czech // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). , 2018.
69. Mohri M. Weighted Finite-State Transducer Algorithms. An Overview // Formal Languages and Applications Studies in Fuzziness and Soft Computing. / под ред. С. Martín-Vide, V. Mitrana, G. Păun. Berlin, Heidelberg: Springer, 2004. С. 551—563.
70. Nigmatulina Y. O. Sound contraction in russian spontaneous speech and its implication for spoken word recognition // New Perspect. Speech Action Proc. 2nd SJUSK. 2013. С. 127—139.
71. Nilsson T. Speech Recognition Software and Vidispine. , 2013.
72. Nkosi M. C. Creation of a pronunciation dictionary for automatic speech recognition : a morphological approach. , 2012.
73. Novak J. R., Minematsu N., Hirose K. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework // Nat. Lang. Eng. 2016. Т. 22. № 6. С. 907—938.
74. Palaz D., Collobert R. Analysis of CNN-based speech recognition system using raw speech as input. : Idiap, 2015.
75. Pan J. и др. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling // 2012 8th International Symposium on Chinese Spoken Language Processing. : IEEE, 2012. С. 301—305.
76. Pereltsvaig A. Languages of the World. : Cambridge University Press, 2020.
77. Povey D. и др. The Kaldi speech recognition toolkit. : IEEE Signal Processing Society, 2011.
78. Rabiner L. R. A tutorial on hidden Markov models and selected applications in speech recognition // Proc. IEEE. 1989. Т. 77. № 2. С. 257—286.

79. Rao K. и др. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). : IEEE, 2015. С. 4225—4229.
80. Reddy и др. The Hearsay-I Speech Understanding System: An Example of the Recognition Process // IEEE Trans. Comput. 1976. Т. C—25. № 4. С. 422—431.
81. Schlippe T. и др. Automatic error recovery for pronunciation dictionaries // Thirteenth Annual Conference of the International Speech Communication Association. , 2012.
82. Skrelin P. и др. CORPRES // Text, Speech and Dialogue Lecture Notes in Computer Science. / под ред. P. Sojka и др. Berlin, Heidelberg: Springer, 2010. С. 392—399.
83. Sloboda T., Waibel A. Dictionary learning for spontaneous speech recognition // Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96. , 1996. С. 2328—2331 т.4.
84. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions // J. R. Stat. Soc. Ser. B Methodol. 1974. Т. 36. № 2. С. 111—133.
85. Strik H., Cucchiarini C. Modeling pronunciation variation for ASR: A survey of the literature // Speech Commun. 1999. Т. 29. № 2. С. 225—246.
86. Taylor P. Hidden Markov models for grapheme to phoneme conversion // Ninth European Conference on Speech Communication and Technology. , 2005.
87. Vaswani A. и др. Attention is all you need // ArXiv Prepr. ArXiv170603762. 2017.
88. Vinyals O., Ravuri S. V., Povey D. Revisiting recurrent neural networks for robust ASR // 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). : IEEE, 2012. С. 4085—4088.
89. Vozila P. и др. Grapheme to phoneme conversion and dictionary verification using graphonemes. , 2003.

90. Yolchuyeva S., Németh G., Gyires-Tóth B. Transformer based grapheme-to-phoneme conversion // ArXiv Prepr. ArXiv200406338. 2020.
91. Zhang C., Woodland P. C. A general artificial neural network extension for HTK // Sixteenth Annual Conference of the International Speech Communication Association. , 2015.

Приложение. Варианты произносительных словарей для распознавания окончаний прилагательных.

1. Идеальная транскрипция

Орфография	Транскрипция
ая	a j a
ая	á j a
ее	ɨ j i
ие	i j i
ие	ί j i
ие	ɨ́ j i
ое	a j i
ое	ó j i
ое	ɨ j i
ую	u j u
ую	ú j u
ые	ɨ j i
ые	ɨ́ j i
юю	u j u
яя	a j a
яя	i j a

2. Транскрипция с наибольшим количеством вариантов

Орфография	Транскрипция
ая	a i
ая	a j a
ая	a j i
ая	á i
ая	á j a
ая	á j i
ее	a e
ее	e j e
ее	ɨ i
ее	ɨ j e
ее	ɨ j i
ие	i i
ие	i j i
ие	ɨ́ i
ие	ɨ́ j i
ие	ɨ i
ие	ɨ j i
ие	ɨ́ j i
ое	a i
ое	a j a
ое	a j e

ое	a j i
ое	ó i
ое	ó j e
ое	ó j i
ое	ɨ j e
ое	ɨ j i
ую	u j u
ую	u u
ую	ú j u
ые	ɨ i
ые	ɨ j e
ые	ɨ j i
ые	ɨ́ i
ые	ɨ́ j i
юю	u j u
яя	a j a
яя	i a
яя	i i
яя	i j a

3. Транскрипция из наиболее часто встречающихся гипотез системы

Орфография	Транскрипция
ая	a
ая	a j
ая	ɨ j
ее	ɨ
ее	j
ие	i
ие	í
ое	a
ое	a j
ое	ɨ j
ое	ó j
ую	u
ые	ɨ́
ые	ɨ j
юю	u
яя	i

4. Транскрипция с ограниченным количеством вариантов

Орфография	Транскрипция
ая	é
ая	i̯ j
ее	i̯ i
ие	i
ие	í
ое	é
ое	i̯ j
ое	ó i
ую	u j
ые	í
ые	i̯ j
юю	u j
яя	i

5. Таблица с транскрипциями без /j/ (кроме как перед /u/)

Орфография	Транскрипция
ая	a i
ая	á i
ее	a e
ее	ɨ e
ее	ɨ i
ие	i i
ие	́i i
ие	ɨ i
ие	́ɨ i
ое	a i
ое	a e
ое	ó i
ое	ó e
ое	ɨ e
ое	ɨ i
ую	u j u
ую	ú j u
ые	ɨ i
ые	ɨ e
ые	́ɨ i
юю	u j u

яя	a i
яя	i a
яя	i i