ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

на тему: Интерпретация моделей распознавания речи
тема на английском: Speech Recognition Interpretation

Студентка 2 курса группы БКЛ-201
Шерман Ксения Валерьевна

Научный руководитель
Сериков Олег Алексеевич
Преподаватель

Москва, 2022 г.

Оглавление

1. Введение

Одним из недавних прорывов в области Искусственного Интеллекта, в том числе и в области компьютерной лингвистики, стало появление моделей, называемых трансформерами. Подобные технологии позволяют решать бесчисленное количество заданий, от распознавания частей речи и до оценок тональности текста и выявления именованных сущностей. Трансформеры используются не только в научных, но и в практических целях. Примером подобных моделей могут послужить “голосовые помощники”. Если раньше они существовали только в виде письменных чат-ботов, то сейчас каждый может воспользоваться функциями голосового ввода. Действия, происходящие внутри модели, позволяют быстро распознавать звучащую речь, производить поиск ответа и реализовать устный ответ.

Интерес вызывает то, что процесс и результат работы трансформеров сложно интерпретируемый. С каждым годом модели получают всё более сложную структуру, которая мало связана с реальностью. Сети, на которых основаны модели, изменяют параметры, подстраиваясь под информацию, однако из структуры невозможно понять, получает ли модели ИИ знания о каких-либо лингвистических явлениях, и если да, то какие именно. Таким образом, говоря о системах распознавания речи, для людей до сих пор остаётся загадкой вопрос о том, что именно эти системы знают о фонетических особенностях языка. Я попыталась узнать, какие знания могут скрываться за множественными слоями модели-трансформер wav2vec-2.0.

В этой работе я представлю возможные методы анализа и интерпретации систем распознавания речи с дальнейшим предоставлением и обсуждением полученных результатов исследования модели wav2vec 2.0.

Задачи, которые предстояло выполнить:

1. Определить области фонетики, которые будут затронуты в исследовании
2. Разработать методы исследования и анализа ‘знаний’ модели wav2vec 2.0
3. Сконструировать тесты для каждой из исследуемых проблем, на основе которых будут оцениваться ‘знания’ модели

В процессе работы wav2vec 2.0 преобразует полученный на вход звук в список векторов, являющимися репрезентациями услышанных в аудиозаписи звуков. Исследование знаний модели строится на изучении данных векторов. Они состоят из 392 чисел. Каждое число отсылает к некоторому звуку или элементу. Элементом

я называю те символы, которые определяются моделью, но не являются звуками: <pad>, <s>, </s> и <unk>. Они отмечают промежутки между звуками либо звуки, которые модели оказались неизвестны. Так, например, первое число в векторе может относиться к звуку [a], второе - к [b] и так далее в каждом векторе. Само число представляет собой уверенность модели в том, что услышанный звук является именно тем, на месте которого оно стоит. Чем оно больше, тем больше модель уверена в том, что полученная на вход звуковая волна отсылает к соответствующему звуку.

```
[ 2.0059953 , -3.545824  , -3.5602784 , -3.306668  ,  1.1123396 ,
 -0.12882468,  2.8014376 ,  0.35074753, -0.23479411,  0.39785522]
```

Первые 10 элементов вектора звука [p], которые соответствую элементам и звукам <pad>, <s>, </s>, <unk>, [n], [s], [t], [ə], [l], [a], [i]. По вектору можно понять, что элемент <pad> имеет более высокую степень уверенности, чем, например, звук [a].

Для оценивания знаний модели я выбрала несколько характеристик звуков: гласность/согласность, мягкость/твёрдость, звонкость/глухость. Были выбраны именно эти характеристики, потому что их можно назвать базовыми в связи с тем, что они наиболее ясно осознаются слушателями. Также я решила более детально изучить вектора на выходе модели, чтобы понять, как модель распределяет степени уверенности звуков: есть ли какая-то связь между тем звуком, который был угадан, и теми, которые также получили высокую степень уверенности. Кроме этого, меня заинтересовала область фонем и аллофонов, поэтому последний из тестов касался вопроса понимания моделью распределения аллофонов на фонемы, более точно, может ли модель wav2vec 2.0 различать звуки одинаковых по звучанию аллофонов разных фонем.

Таким образом, мною были сформулированы следующие гипотезы:

1. Модель wav2vec 2.0 понимает деление звуков на гласные и согласные, а также различает звуки по глухости и звонкости, мягкости и твёрдости.
2. Модель wav2vec 2.0 понимает, что один звук может являться аллофоном различных фонем и может определить, аллофон какой фонемы представлен в слове.
3. Модель wav2vec 2.0 знает, какие звуки похожи по своей реализации.

Исследование проводилось с использованием языка программирования Python. Код лежит в репозитории по следующей ссылке: https://github.com/Supermiledi/speech-recognition-interpretation

2. Обзор литературы

2.1. Модель

Изучаемая модель wav2vec 2.0 [Baevski et al. 2020] была создана в 2020 году и относится к моделям-трансформерам. Она состоит из трёх блоков: кодировщик признаков, преобразователь и модуль квантования. Кодировщик состоит из многослойной сверточной нейронной сети, которая преобразует необработанное входное аудио в новое представление, называемое скрытым. Преобразователь принимает на вход скрытые вектора и строит новые представления, которые используются для обучения. Обучение модели осуществляется путем маскирования определённого промежутка времени (некоторого количества чисел в векторе) в представлении скрытых признаков и дальнейшего применения контрастивной задачи. Цель этой задачи - найти правильное квантованное представление, соответствующее замаскированному звуковому представлению среди набора отвлекающих факторов. Соответствующие квантованные представления получаются в результате передачи вектора, полученного из кодировщика, в блок модуля квантования, предварительно дискретизировав, чтобы получить конечный набор представлений речи. Схема модели wav2vec 2.0 представлена на рисунке 1.

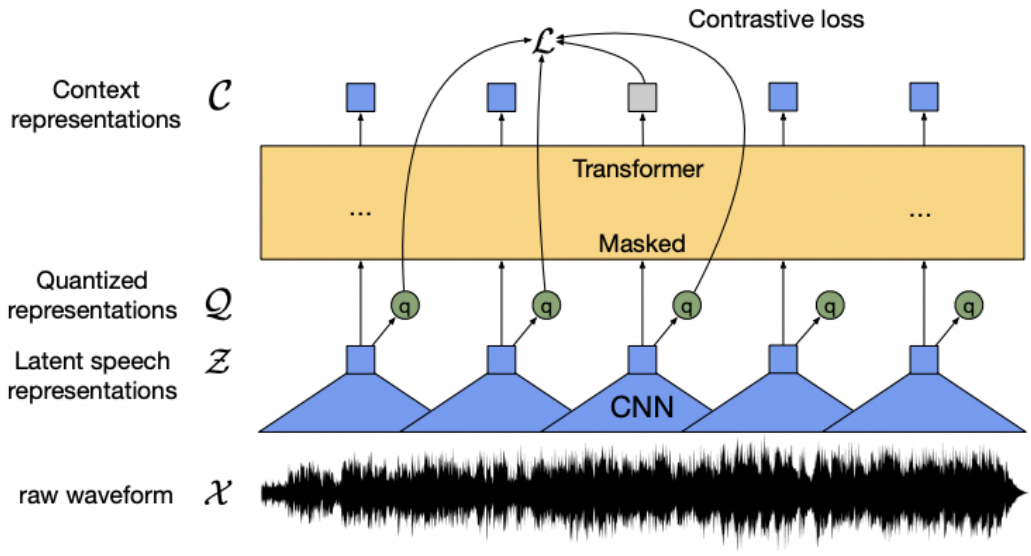


Рисунок 1. Схематичная иллюстрация структуры модели wav2vec 2.0

Модель wav2vec 2.0 была предобучена на неразмеченных данных датасета LibriSpeech, а после отрегулирована на размеченных данных датасета TIMIT, который используется для решения задачи определения услышанных звуков. Результат обучения модели оценивается метрикой WER (word error rate), который считает долю ошибок, сделанных моделью. Таким образом, модель достигает 1.8 и 3.3 WER на чистых данных и данных с шумом соответственно при обучении на всех данных датасета Librispeech.

2.2. Пробинг

При изучении векторов, которые получаются в результате работы сложных моделей, используется задачи, которые называются *пробинг*. Задание пробинга — это задание классификации данных на основе простых лингвистических свойств языкового элемента [Conneau et al. 2018]. Система пробинга описывается 3 принципами:

1) для решения задачи пробинга ставится простой вопрос, что позволяет сделать интерпретацию результатов более понятной;

2) Из-за простоты легче контролировать предвзятость модели, чем при обычных задачах (например, тональность текста);

3) Тесты пробинга не зависят от архитектуры кодировщика.

В качестве модели пробинга может использоваться простая модель классификации - логистическая регрессия. В более сложных работах создаются собственные модели или нейронные сети, как, например, в [Tenney et al. 2019].

Результаты модели оцениваются с помощью метрики accuracy. Тем не менее, одних результатов метрики оказывается недостаточно. Существует вероятность, что высокие результаты работы классификатора были получены случайно. Для проверки этого явления были предложены контрольные задачи [Hewitt, Liang 2019]. Идея таких задач основана на интуиции, согласно которой чем лучше результаты предсказаний модели независимо от лингвистических свойств представления некоторого явления, тем меньше точность, полученная в процессе обучения модели, отражает свойства вектора в лингвистической задаче. Так, модель описывается как *selective*, если получает высокие значения accuracy при оценке лингвистической задачи и низкие – при оценке контрольной.

Лингвистические модели получают много внимания исследователей, но в основном это модели, которые обрабатывают и анализируют письменную речь. Наиболее известной на данный момент можно назвать модель BERT [Devlin et al. 2019]. Множество исследований было проведено на предмет знания моделей о языке в области морфологии, семантики и синтаксиса. Так, было выяснено, что вектора BERT хранят информацию о части речи и синтаксической роли, а также было показано, что модель кодирует информацию и о семантических ролях слов и типах сущностей [Tenney et al. 2019]. Часто производятся сравнения результатов на разных скрытых слоях модели, например, было обнаружено, что на различных слоях модель BERT производит вектора с разной степенью опоры на контекст: на верхних слоях вектора сильнее отражают контекст слова, чем на нижних [Wiedemann et al. 2019].

До появления моделей-трансформеров для решения задачи распознавания речи использовались нейронные сети, например, DNN (Deep Neural Networks) - глубокие нейронные сети [Nagamine et al. 2015], предметом исследования которых были отдельные узлы нейронной сети на разных слоях и то, за какие характеристики звука каждый из узлов отвечает. Различные фонетические особенности были исследованы в работах [Alishahi et al. 2017; Belinkov, Glass 2017], где внимание уделялось тому, как точно модели автоматического распознавания речи определяют различные звуки, например, по способу их образования. Помимо этого, подобные модели были исследованы на возможность собирать информацию об акценте и о характеристиках говорящего [Prasad, Jyothi 2020; Raj et al. 2019]

По модели wav2vec было написано небольшое количество работ. В одной из них [Ma et al. 2021] авторы сравнили несколько моделей распознавания речи с помощью предложенных ими методов пробинга, выяснив в результате, что модели, действительно, содержат в себе информацию о различных фонетических особенностях звуков, таких как определение согласного или гласного звука, определение его фрикативности, а также предугадывание звуков. Есть работы [Shah et al. 2021], в которых исследования проводились уже с моделью wav2vec 2.0. В результате изучения авторы определили, что модель сохраняет некоторую информацию об особенностях произношения различных звуков, например, пол говорящего и его акцент. Однако ещё не было проведено исследований, которые бы позволили понять,

кодирует ли каким-либо образом информацию о фонетических характеристиках звуков модель wav2vec 2.0.

3. Методы

Один из тестов, который часто используется для того, чтобы оценить знания модели ИИ, это обучение классификатора на полученных из выхода модели векторах, в том числе с каждого из скрытых слоёв, и исследуемых данных. В этой работе я буду использовать классификатор Logistic Regression из библиотеки sklearn. Для простоты повествования модель wav2vec 2.0 в дальнейшем я буду называть модель Wav2Vec, а классификатор - модель LogReg.

Модель LogReg принимает на вход список векторов чисел и распределение классов. Один вектор соответствует одному объекту. Каждый вектор построен таким образом, чтобы числа на одних и тех же местах относились к одному признаку для каждого объекта. Например, в векторах слов первое число может описывать количество вхождений данного слова в первый текст, второе – во второй и т.д. Сами числа называются признаками. Модель LogReg обучается предсказывать вероятность, что выбранный объект принадлежит некоторому классу. В процессе обучения линейная регрессия присваивает веса подаваемым на вход признакам, называемых также коэффициентами. Результатом скалярного произведения этого вектора коэффициентов с вектором звука является число, которое позволяет определить класс, к которому относится объект соответствующего вектора. Сами коэффициенты, позволяют определить, какие из признаков модель LogReg считает наиболее важными для предсказания. Чем выше вес у элемента вектора, тем более важным для предсказания его считает модель. В связи с этим появляется одно из предположений, которое говорит о том, что если выбор весов можно объяснить с помощью реальных фонетических правил и особенностей звуков, то модель Wav2Vec владеет некоторой информацией об исследуемом явлении, а также о сходстве звуков.

Вектора, полученные на выходе модели Wav2Vec подаются на вход модели LogReg для обучения и предсказывания. Правильность предсказывания оценивается с помощью метрики accuracy — доля правильно угаданных ответов. Для исследования высокое значение метрики показывает, что распределение информации в векторах похоже на распределение информации в данных. Также следует ввести

контрольную задачу. В данной работе я решила проверять полученные результаты с помощью тестов двух видом перемешивания:

1) перемешать полностью все результаты, не обращая внимания на вектора одинаковых звуков. Так, два вектора, которые соответствуют одному звуку, могут получить разные результаты - один будет гласным, другой - согласным;

2) выбрать некоторые звуки и всем векторам этих звуков поменять результат. Так, например, все вектора звука [a] могут оказаться согласными.

В дальнейшем первый способ я буду называть полным перемешиванием, второй - групповым перемешиванием.

4. Данные

Для оценки знаний модели были использованы 2 набора данных (датасета):

1) common voice [Shah et al. 2020], доступный на ресурсе hugging face, для тестов на гласные/согласные, звонкие/глухие, мягкие/твёрдые и для исследования векторов отдельно;
2) TIMIT [Garofolo et al. 1993] для тестов на фонемы

Первый датасет содержит более 9000 часов аудиозаписей на 60 языках. Для экспериментов всех экспериментов использовались записи русской речи. К каждой записи прикладываются данные о том, какое предложение было произнесено, а также характерные черты говорящего. В датасете отсутствует транскрипция записи и разделение самой записи на звуки или слова, поэтому не было возможности проверить точность предсказания модели и тот факт, что полученный вектор, действительно, принадлежит предсказанному звуку. Тем не менее, результаты тестирования модели Wav2Vec позволяют сделать допущение о том, что полученные данные достаточно точны, чтобы проводить исследование, но вероятность неправильной оценки услышанного звука существует, хотя и незначительная. В связи с ограниченными возможностями компьютера, результаты первых четырёх тестов были получены на сравнительно небольшом количестве данных - 300 аудиозаписей.

Датасет TIMIT содержит 6300 аудиозаписей на английском языке, которые были записаны 630 говорящими. Так же, как и в common voice, для каждой записи известны метаданные говорящего. Однако в отличие от первого датасета, в TIMIT есть разбиение каждой записи на слова и на звуки, для которых известны начало и

конец реализации, что позволяет регулировать контексты, в которых возникают звуки, для более точных тестов, что требовалось для тестов на фонемы. Датасет описан в транскрипции, не опирающейся на МФА.

5. Фонетические особенности

В первой секции я детально анализирую вектора, порождённые моделью в процессе распознавания датасетов, чтобы оценить знания модели о сходстве звуков по способу их реализации. Следующие 3 секции этой главы посвящены тестам ‘гласный/согласный’, ‘звонкий/глухой’ и ‘мягкий/твёрдый’. В последней секции я разбираю результаты тестов на аллофоны и фонемы.

5.1. Сходство звуков

Основное предположение состоит в том, что в векторе каждого из звуков может быть закодирована информация о свойствах, которыми он обладает, и следовательно, есть возможность определить похожие друг на друга звуки по тем или иным параметрам: например, можно назвать звуки [p] и [b] похожими по способу и месту реализации с отличием в участии голосовых связок. Каждый вектор является набором чисел, представляющих уверенность модели Wav2Vec, где каждому индексу сопоставлен определённый звук. Тогда, для поиска похожих звуков требовалось оценить значения, представленные в векторе. В связи с тем, что реализации звука отличаются в разных словах, потребовалось взять все вектора и усреднить значения. Пример трёх векторов можно посмотреть на рисунке 2. Графическое представление того, какие звуки модель считает наиболее похожими можно посмотреть в приложении.

При детальном изучении каждого из векторов, удалось заметить несколько интересных моментов. Действительно, у всех векторов большие коэффициенты получают звуки, похожие по реализации на угаданный. Так, у векторов гласных зву-

```
['p', 'b', 'k', 't', '<pad>', 'pʲ', 'f', 'm', 'v']
['b', 'v', 'p', 'd', 'ɡ', 'm', 'β', 'bʲ', 'ʊ']
['a', 'ʌ', 'ɑ', 'aː', 'ɑː', 'ɐ', 'ɛ', 'ɔ', 'e']
```

Рисунок 2. Первые 9 элементов с наибольшими степенями уверенности в усреднённых векторах звуков [p] (верхний), [b] (средний) и [a] (нижний)

ков высокие степени уверенности получили гласные звуки, а у согласных - согласные. Для гласных часто оказываются наиболее близкими звуки, несильно отличающиеся по подъёму или ряду. Так в векторе звука [ɔ] высокую степень уверенности получил звук [o], а в векторе звука [ɑ] - звук [a]. В векторах звуков [o] и [u] наиболее близкие звуки оказались огубленные. Тем не менее, сложно определить какую-то иерархию, относительно которой работает распределение степеней уверенности модели.

Аналогичным образом можно проверить вектора согласных звуков. В векторах многих парных согласных, например, [t] и [d] или [k] и [g] высокие значения уверенности получают соответствующие им парные звуки. У звуков назального способа образования наиболее близкими оказываются также носовые согласные, например в векторе [m] большое значение получает звук [n]. Многие фрикативные звуки получили высокие значения у элемента <pad>, который используется для обозначения временного периода, в котором нет реализации звука. В векторах мягких звуков большую степень уверенности получили другие мягкие звуки, когда как в векторах твёрдых они не встречались на высоких позициях.

Однако были и необычные случаи. Например, модель дала высокую степень уверенности звуку [r] в векторе звука [h], хотя оба звука отличаются как по месту, так и по способу образования. Также в векторе звука [ð] высокое значение получили звуки [l] и [b]. В дополнение к этому, в векторе звука [ɾ] высокие положения заняли некоторые гласные: [ə], [ɛ], [e], [a].

Данные наблюдения позволяют сделать вывод о том, что модель имеет некоторые знания о свойствах звуков, так как умеет группировать их по сходству в реализации. Несмотря на некоторые неточности, которые были описаны в предыдущем абзаце, можно сказать, что высокие степени уверенности, действительно, получают те звуки, реализация которых похожа на реализацию звука, относящегося к соответствующему вектору.

5.2. Гласный или согласный звук

Тест на гласный или согласный звук показал высокие значения accuracy на каждом из слоёв с увеличением результата с последующими слоями, достигнув своего максимума в значении 0.99 на последних слоях. Полное перемешивание показывает невысокую точность предсказания - около 0.63, которая и ожидается. Групповое перемешивание, несмотря на неточность подающихся данных, даёт очень высокий результат - 0.98-0.99 на последних слоях. Изменения значений accuracy можно увидеть на графике 1. Можно заметить, что selectivity, то есть разница между значениями метрики на верных данных и перемешанных, постепенно сокращается, следовательно, первые слои трансформера лучше определяют гласные и согласные звуки, когда как при обучении на векторах с последних слоёв модель LogReg, скорее всего, просто запоминает данные.

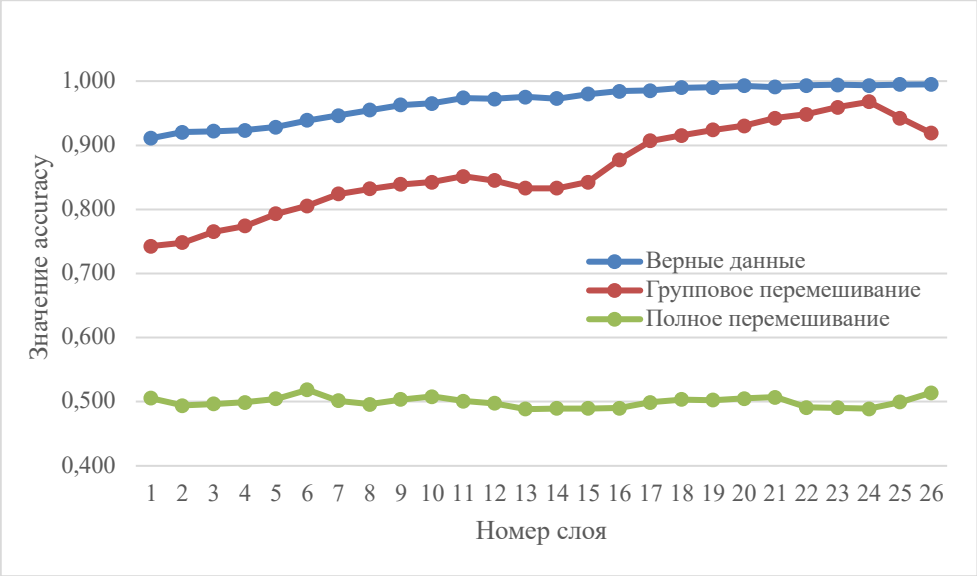


График 1. Изменения значений accuracy модели LogReg на векторах разных слоёв модели wav2vec 2.0 при верных данных и перемешанных на тесте ‘гласный/согласный’.

Коэффициенты, которые модель LogReg дала элементам вектора, оказались трудно объяснимыми. Основываясь на идее, что чем больше по модулю коэффициент логистической регрессии, тем более важен для предсказания признак, я предполагала, что модель разделит коэффициенты таким образом, что звуки одного класса получат коэффициенты одного знака. Однако получилось так, что среди звуков с весами одного знака встречаются звуки разных классов. Так, среди гласных,

которым модель LogReg определила большие положительные веса, оказался согласный звук [ɹ], а среди отрицательных – гласные [y] и назализованный [o] (график 2).

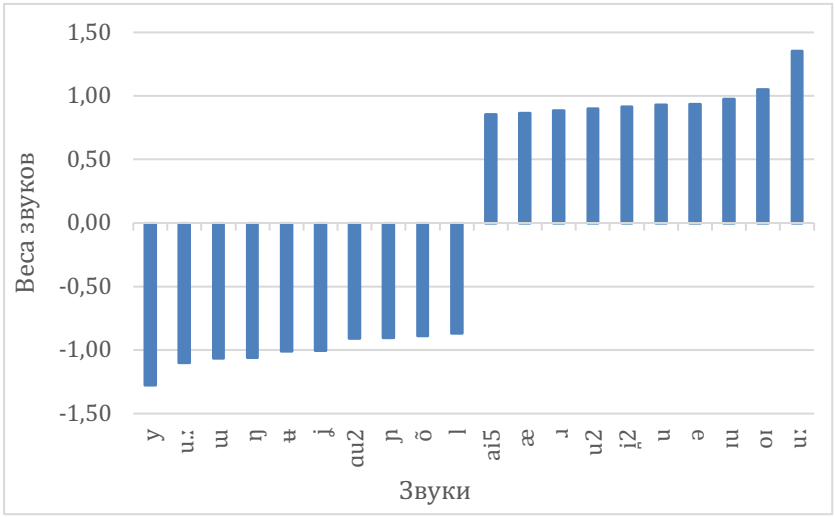


График 2. Звуки с наибольшими по модулю коэффициентами логистической регрессии на тесте ‘гласный/согласный’.

На первый взгляд кажется, что вектора модели wav2vec случайны, так как модели LogReg удалось обучиться на перемешанных данных. Тем не менее, низкие результаты тестов на полностью перемешанных данных позволяют сделать вывод, что все вектора, соответствующие одному звуку, построены похожим образом. Другими словами, вектора, например, звука [j] настолько похожи между собой, что не могут относиться к разным классам, как предполагает тест с полным перемешиванием. Следовательно, сами вектора были получены не случайно, и модель Wav2Vec имеет некоторые знания о фонетике. Так, в векторах гласных звуков высокая степень уверенности будет также у гласных звуков. Однако из-за факта, что модели LogReg удалось хорошо обучиться при групповом перемешивании данных, нельзя сказать, что информация о том, какой звук соответствует вектору, гласный или согласный, закодирована в векторе.

5.3. Звонкость

Тест на звонкость показывает высокое значение accuracy: доля правильно предсказанных звуков стремится к значению 0,99 на последнем слое. Тесты на перемешивание показали похожие результаты с тестами по гласному/согласному звуку: при групповом перемешивании модель на последних слоях выдала резуль-

тат, почти равный результату на обычных данных, - 0,98 а при полном перемешивании 0.629 (график 3). Аналогично предыдущему тесту, оказывается, что selectivity модели уменьшается в процессе перехода к более глубоким слоям модели. Получается, что вектора на первых слоях содержат больше информации о звонкости звука, чем на последних.

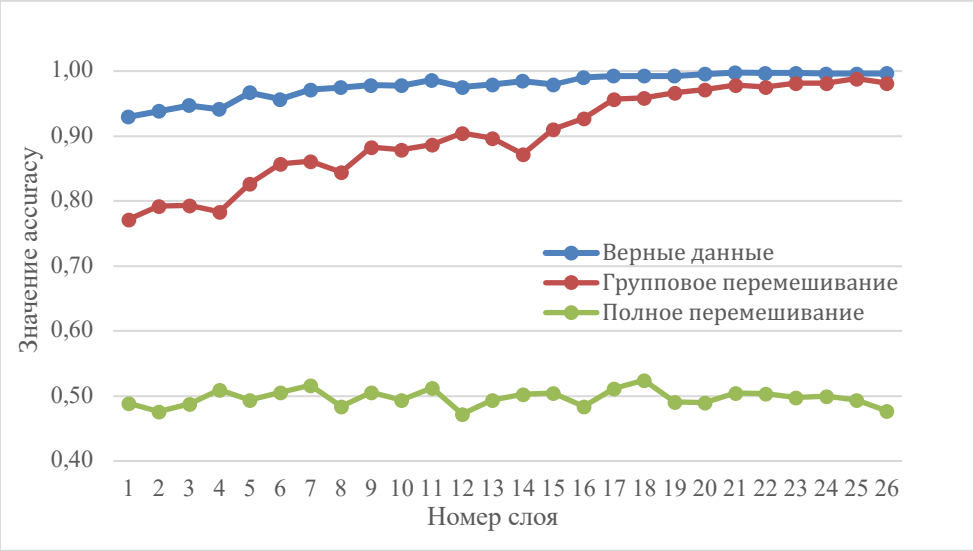


График 3. Изменения значений accuracy модели LogReg на векторах разных слоёв модели wav2vec 2.0 при верных данных и перемешанных на тесте ‘звонкий/глухой’.

Если обратить внимание на коэффициенты, которые определила модель LogReg при обучении, то можно заметить, что наибольшие коэффициенты она дала глухим звукам [t], [h], [s], а наименьшие - звонким [z], [r], [m] и т.д. (график 4). Таким образом, можно сказать, что при обучении данные звуки оказываются наиболее показательными для определения звонкости согласного, следовательно,

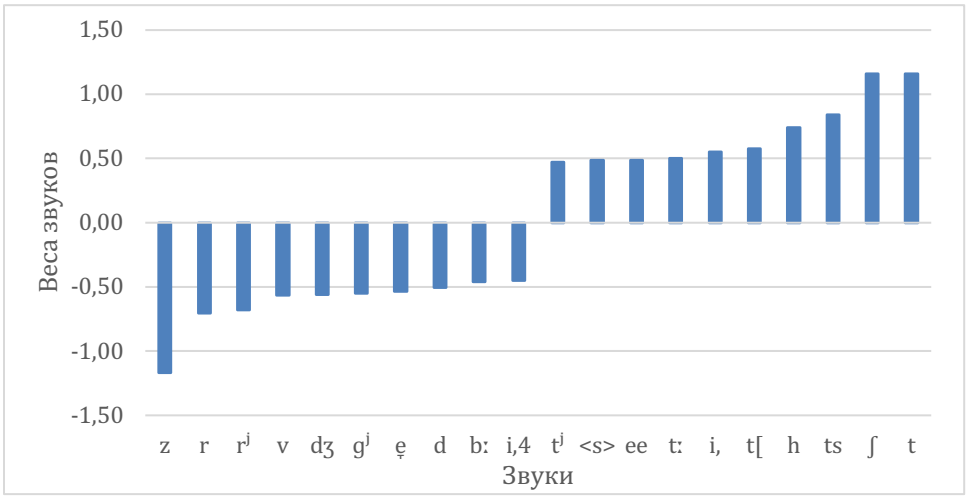


График 4. Звуки с наибольшими по модулю коэффициентами логистической регрессии на тесте ‘звонкость/глухость.

у всех глухих звуков будет большое значение у числа, соответствующего звука, например, [t]. Интересным оказывается тот факт, что модель LogReg дала высокий вес гласному звуку [i], хотя гласные должны быть скорее больше похожи на сонорные звуки, так как тоже используют связки при реализации, что нельзя сказать про глухие звуки (график 4).

Таким образом, аналогично предыдущему тесту, нельзя утверждать, что модель эксплицитно кодирует информацию о сонорности звука, однако вектора построены не случайным образом - в векторе глухого звука значения чисел, которые отсылают к другим глухим звукам, будут выше, чем значения при звонких звуках, и наоборот.

5.4. Палатальность

Результат предсказаний мягкости и твёрдости моделью растёт с каждым последующим слоем, достигая максимума 0,96. Тем не менее оказалось необычным то, что модель показывает более хороший результат при групповом перемешивании: превысив по значениям результаты теста на верных данных на 21 слое, значения при групповом перемешивании достигают 0, 97 на последнем слое. Значения accuracy при полном перемешивании не превышают 0,54 (график 5). В отличие от предыдущих тестов, получившиеся результаты указывают на низкий selectivity на всех слоях модели. Тем не менее, тенденция остаётся такой же: на первых слоях трансформера свойства мягкости звука определяются лучше.

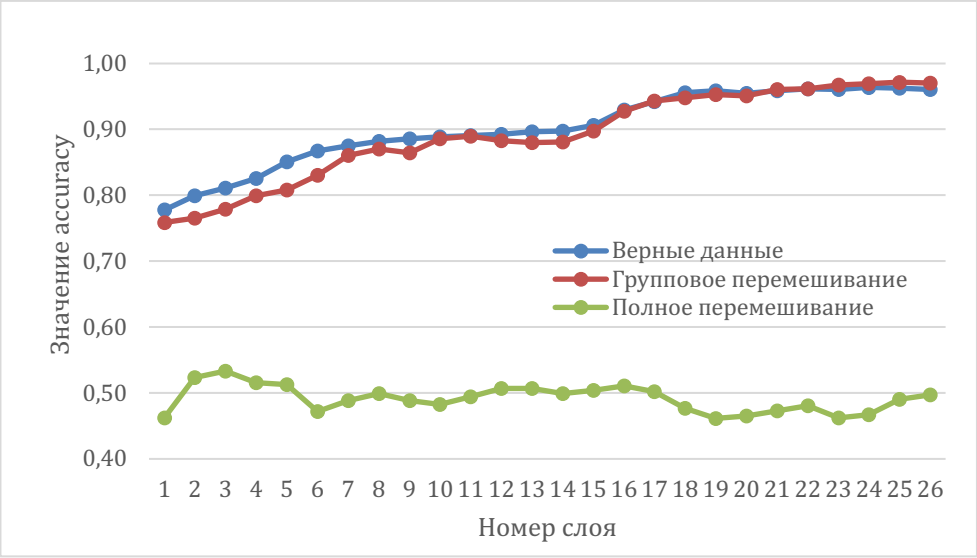


График 5. Изменения значений accuracy модели LogReg на векторах разных слоёв модели wav2vec 2.0 при верных данных и перемешанных на тесте ‘мягкий/твёрдый.

Этот случай превышения значения accuracy при групповом перемешивании оказался единственным, поэтому было интересно понять причину более низких результатов на верных данных. Следовало рассмотреть те случаи, в которых Wav2Vec и LogReg давали разные результаты. В основном различия в ответах были тогда, когда модель Wav2Vec предсказывала твёрдый звук, а логистическая регрессия - мягкий. Оказалось, что многие звуки, которые модель Wav2Vec решила определить как твёрдые, должны показывать признаки мягкости, так как часто предшествуют смягчающему гласному [i]. Таким образом, результат, полученный по исходным данным, оказался ниже из-за неточности языковой модели, в то время как модель LogReg оказалась способна по вектору звука предсказать его мягкость.

При обучении модели LogReg на верных данных, звук [j] получил наибольший коэффициент, что естественно, так как это палатальный звук, к которому стремится артикуляция при произнесении более мягких звуков. Высокие веса также получили палатализованные звуки [mʲ] и [ɻʲ].

В процессе исследования выяснилось, что модель Wav2Vec услышала в аудиозаписях как палатализованный [tʃʲ], так и велярный [tʃ], хотя в русском языке выделяется только мягкая реализация этого звука с некоторыми исключениями. Из 300 аудиозаписей, модель услышала 2 звука [tʃʲ] и 145 звуков [tʃ]. В дополнение к этому, модель wav2vec выделила только следующие мягкие звуки [dʲ], [j], [mʲ], [nʲ], [rʲ], [sʲ], [tʲ], [ɻʲ], однако в русском языке встречаются также [pʲ], [kʲ], [bʲ] и другие.

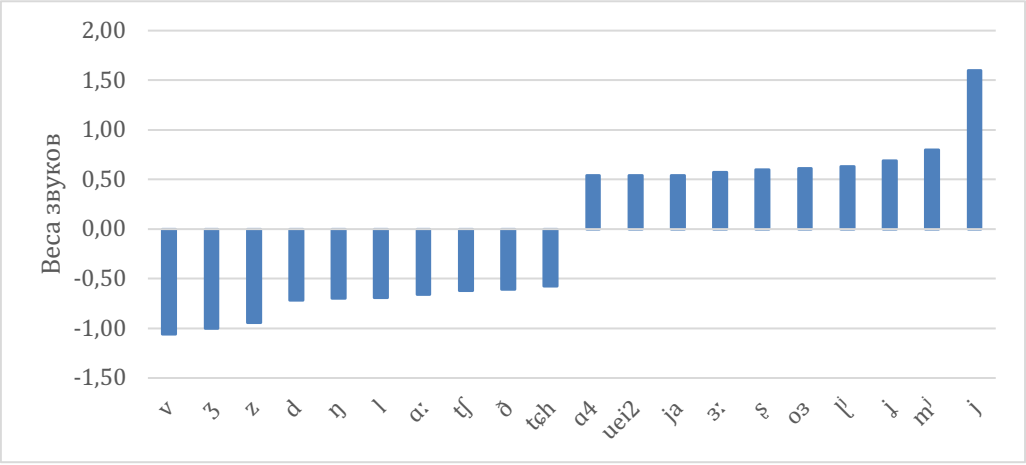


График 6. Звуки с наибольшими по модулю коэффициентами логистической регрессии на тесте ‘мягкость/твёрдость.

Таким образом, способности модели Wav2Vec определять звуки по мягкости не очевидны и требуют более подробного изучения. Так, остаётся неизвестным тот критерий, по которому модель считает звуки мягкими или твёрдыми, так как оказалось, что те реализации, которые должны считаться мягкими в действительности, модель распознавания речи отнесла к твёрдым. Чтобы выяснить причину подобного поведения, следует сделать более подробные тесты на основании различных стимулов, которые влияют на мягкость звука.

5.5. Фонемы

В глубинном представлении каждого из слов встречается то, что называется фонемами - абстракция, которая лежит в основе каждого из звуков и на которую накладываются правила, в результате использования которых возникает определённая реализация звука. Благодаря таким абстракциям мы способны выводить обобщённые правила, позволяющие показать, например, почему в слове код [кот] последний согласный звук [т], а во множественном числе [коды] тот же звук сменяется на [д]. Один и тот же звук может оказаться аллофоном нескольких фонем.

Возникает вопрос о том, способна ли модель Wav2Vec определять, являются ли две реализации аллофонами одной фонемы или разных. Для точной оценки этих данных понадобилось провести два теста на две различные пары фонем с одинаковыми аллофонами: /s/ и /z/, /t/ и /d/. Для первого теста было необходимо оценить вектора получившихся звуков. Предполагалось, что вектора звуков, которые являются аллофонами различных фонем, будут отличаться. Для второго теста требовалось обучить модель LogReg определять, какая фонема стоит за звуком.

Для теста по фонемам /z/ и /s/ я решила выбрать все существительные, которые оканчиваются на звук [s] или [z]. Дальше я разделила слова на 2 группы. К первой группе я отнесла существительные во множественном числе, так как они оканчиваются на звук, который является аллофоном фонемы /z/. Остальные существительные образовали вторую группу слов, которые оканчиваются на аллофон,

```
['s', '<pad>', 'z', 't', 'ʃ', 'ts', 'θ', 'ə', 'f']
['s', 'z', '<pad>', 'ts', 't', 'ʃ', 'd', 'θ', 'ə']
```

Рисунок 3. Первые 9 элементов с наибольшими степенями уверенности для усреднённого вектора аллофонов фонемы /s/ (сверху) и фонемы /z/ (снизу)

принадлежащий фонеме /s/. Так как количество доступных мне данных были ограничены, подходящих для тестов слов оказалось немного – 248.

Результаты первого теста не показали желаемых результатов. Вектора мало отличались друг от друга и скорее создавались на основании сходства звуков, никак не отражая фонемное распределение (рисунок 3). Обучение модели LogReg показало высокие результаты: значения accuracy колебались от 0.66 на первом слое до 0.88 на предпоследнем (график 7). Однако распределение классов было слишком разнородным: существительных множественного числа оказалось 199, а остальных слов - 49, следовательно, обучение не может считаться полностью корректным. Тест на перемешанных данных показал низкие результаты по сравнению с результатами на верных данных.

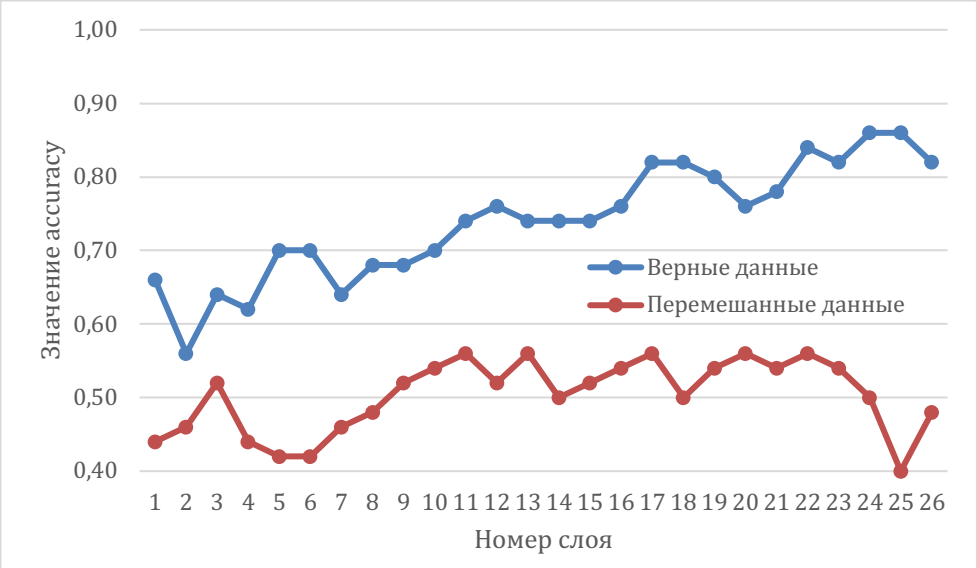


График 7. Изменения значений accuracy модели LogReg на векторах разных слоёв модели wav2vec 2.0 при верных данных и перемешанных на тесте фонем /z/ и /s/.

Данных на сравнение фонем /t/ и /d/ оказалось больше - 1412, с последующим разбиением классов: 370 слов, имеющих окончание прошедшего времени -ed, а следовательно, оканчивающиеся на аллофон фонемы /d/, и 1042 слова, содержащих звук [t] - аллофон фонемы /t/. Сравнение векторов показало, что сильных различий в списке ‘похожих’ звуков не оказалось (рисунок 4), аналогично сравнению

```
['t', '<pad>', 'd', 'ɾ', 'k', 's', 'tʃ', 'p', 'ts']
['t', '<pad>', 'd', 'ɾ', 'k', 's', 'tʃ', 'n', 't:']
```

Рисунок 4. Первые 9 элементов с наибольшими степенями уверенности для усреднённого вектора аллофонов фонемы /t/ (сверху) и фонемы /d/ (снизу)

векторов аллофонов фонем /z/ и /s/. Обучение модели LogReg показало высокие результаты - наибольшее значение 0.92 accuracy на 18 слое, при максимальном значение accuracy при перемешивании равном 0.52 (график 8).

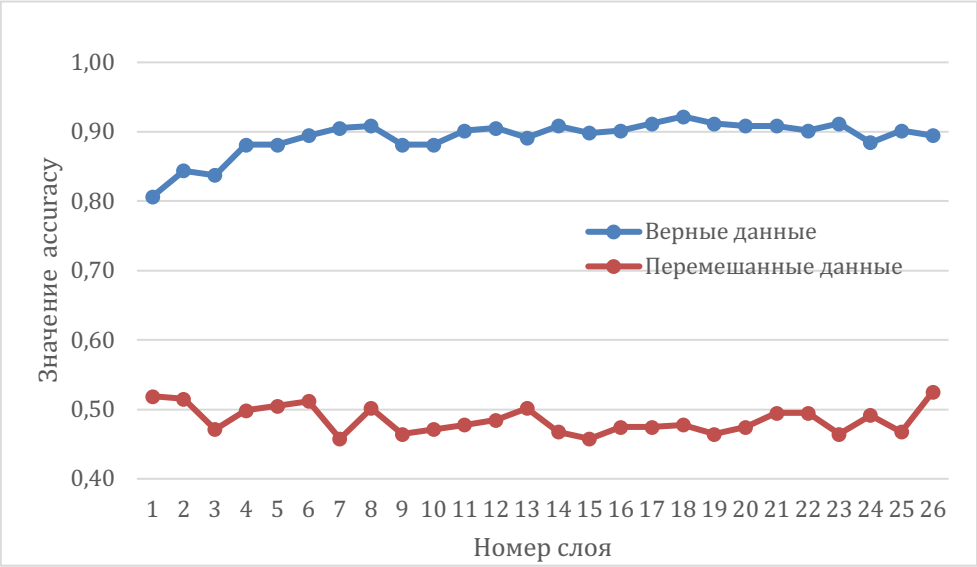


График 8. Изменения значений accuracy модели LogReg на векторах разных слоёв модели wav2vec 2.0 при верных данных и перемешанных на тесте фонем /t/ и /d/.

Однако, если обратить внимание на веса, которые выставила модель LogReg в тестах на каждую пару фонем, окажется, что она считает наиболее важными для оценки некоторые гласные звуки, что, как кажется, не поддаётся объяснению и не соотносится с внутренним пониманием различия аллофонов разных фонем. В том числе, после проверки степени уверенности у звуков, которым модель LogReg выставила наибольшие по модулю коэффициенты, оказалось, что соответствующие значения отличаются между векторами меньше чем на 0,1, то есть выбранные моделью логистической регрессии звуки не являются показательными.

Таким образом, результаты тестов оказались противоречивы. Вектора разных аллофонов мало отличаются и, кажется, не содержат информацию о фонеме. Однако модели LogReg удаётся добиться высоких результатов accuracy, что означает, что по векторам можно научиться определять к какой фонеме относится выбранный звук. Лучше всего фонемы /s/ или /z/ предсказываются по векторам 25 слоя, а

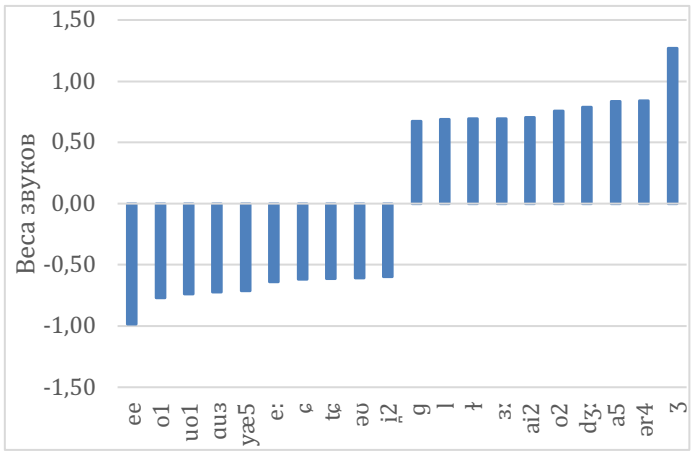


График 8. Звуки с наибольшими по модулю коэффициентами логистической регрессии на тесте фонем /t/ и /d/.

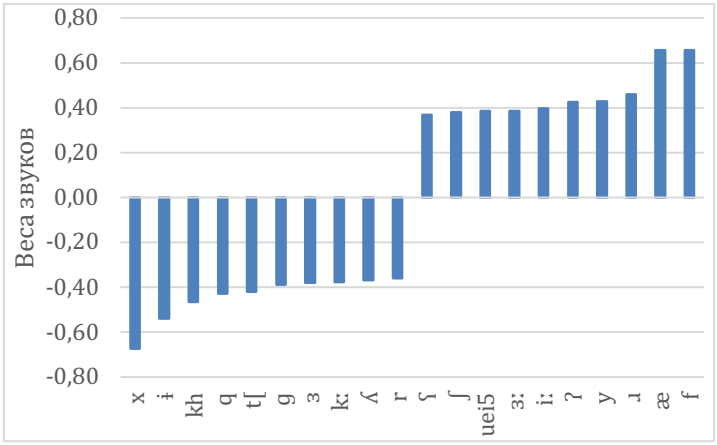


График 9. Звуки с наибольшими по модулю коэффициентами логистической регрессии на тесте фонем /s/ и /z/.

фонемы /t/ и /d/ - 18 слоя. Тем не менее, звуки, которые модель выбрала как наиболее важные для предсказания, имеют схожие значения степени уверенности в обоих векторах.

Вопрос понимания моделью Wav2Vec фонем и аллофонов остаётся открытым. Возможно, на результаты модели LogReg повлияло строгое разделение аллофонов, которое было мною предложено. Так, вместо предсказывания фонемы модель LogReg могла предсказывать множественное число существительного или прошедшее время глагола, так как только эти характеристики были использованы для примеров фонем /z/ и /d/ соответственно. Следовательно, для более точных тестов, следует подобрать больше возможных контекстов для выбранных пар фонем.

6. Заключение

В ходе данной работы было проведено несколько тестов, которые позволяют определить, кодирует ли каким-либо образом модель wav2vec 2.0 информацию об основных фонетических характеристиках звуков. В том числе были более подробно исследованы вектора каждого из звуков, чтобы определить логику распределения чисел внутри вектора.

Таким образом, были выявлены несколько важных особенностей модели wav2vec 2.0. В первую очередь, нельзя утверждать, что модель wav2vec 2.0 не кодирует эксплицитно информацию о том, является ли звук гласным или согласным, глухим или звонким, мягким или твёрдым, следовательно, первая гипотеза не была подтверждена.

Вторая гипотеза также не подтвердилась, нельзя с уверенностью говорить о том, насколько точно модель Wav2Vec различает аллофоны разных фонем. Для поиска верного ответа следует сделать повторные тесты на большем количестве данных с использованием большего количества контекстов.

Третья гипотеза подтвердилась. Действительно, почти для каждого вектора распределение чисел имеет логичное объяснение, следовательно, модель wav2vec 2.0 имеет представление о том, какие звуки наиболее похожи по своей реализации.

Тем не менее, при проведении тестов на большем количестве данных, а также при использовании других инструментов пробинга, можно ожидать более точных результатов на каждом из тестов.

Интерес для дальнейшей работы представляет более глубокое изучение того, как модель определяет мягкие и твёрдые звуки. Например, можно записать различные слова-стимулы, в которых мягкость будет проявляться в различных контекстах, что позволит выяснить, какую из реализаций модель wav2vec 2.0 определяет как мягкую. В том числе, в процессе работы было замечено, что словари услышанных звуков отличались на русском и английском языках. Таким образом, будущие исследования в этой области могут опираться на вопросы о том, как модель слышит разные языки, а также есть ли отличия между векторами одинаково угаданных моделью звуков различных языков.

Литература

Распознавание речи

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // arXiv preprint arXiv:2006.11477. - 2020.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue. Timit Acoustic-Phonetic Continuous Speech Corpus. // Web Download. Philadelphia: Linguistic Data Consortium. - 1993.

Jui Shah, Yaman Kumar Singla, Changyou Chen, Rajiv Ratn Shah. Common Voice: A Massively-Multilingual Speech Corpus // arXiv preprint arXiv:2101.00387. - 2020.

Пробинг

Afra Alishahi, Marie Barking, Grzegorz Chrupała. Encoding of phonology in a recurrent neural model of grounded speech // arXiv preprint arXiv:1706.03815. - 2017.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*)*, Melbourne, Australia, July, 2018. P. 2126–2136.

Archiki Prasad and Preethi Jyothi. How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July, 2020. P. 3739–3753.

Danni Ma, Neville Ryant, Mark Liberman. Probing Acoustic Representations for Phonetic Properties // arXiv preprint arXiv:2010.13007. - 2021.

Desh Raj, David Snyder, Daniel Povey, Sanjeev Khudanpur. Probing the Information Encoded in X-vectors // arXiv preprint arXiv:1909.06351. - 2019.

Gregor Wiedemann , SteffenRemus, AviChawla, and ChrisBiemann. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. // arXiv preprint arXiv:1909.10430. - 2019

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie

Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations // arXiv preprint arXiv:1905.06316. - 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. - 2019.

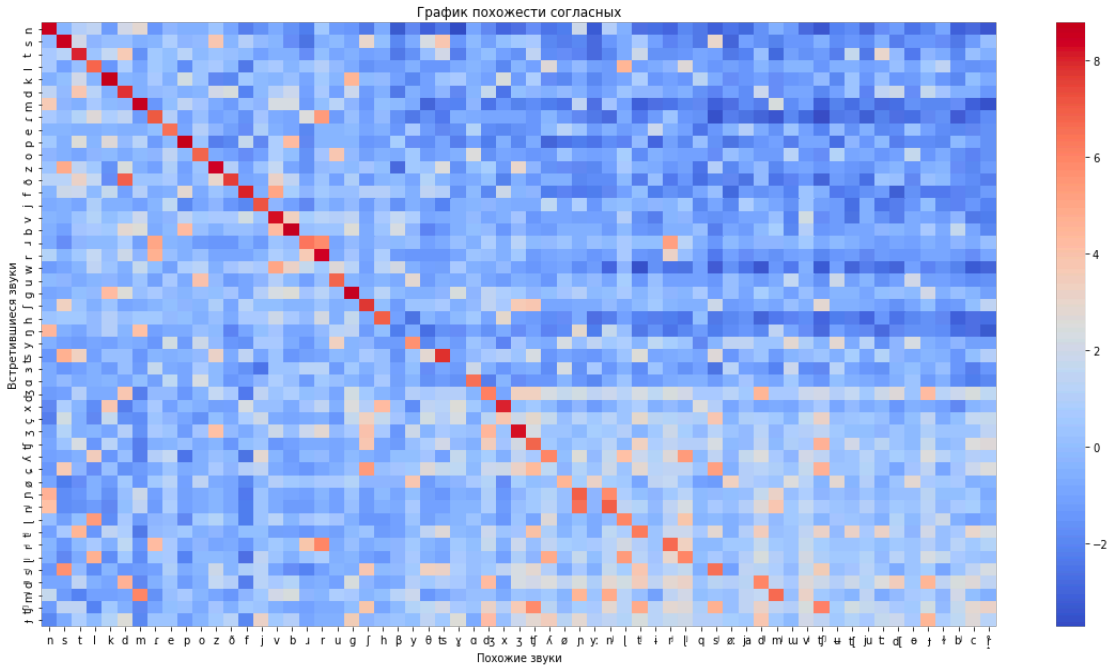
John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November, 2019. P. 2733–2743.

Jui Shah, Yaman Kumar Singla, Changyou Chen, Rajiv Ratn Shah. What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure // arXiv preprint arXiv:2101.00387. - 2021.

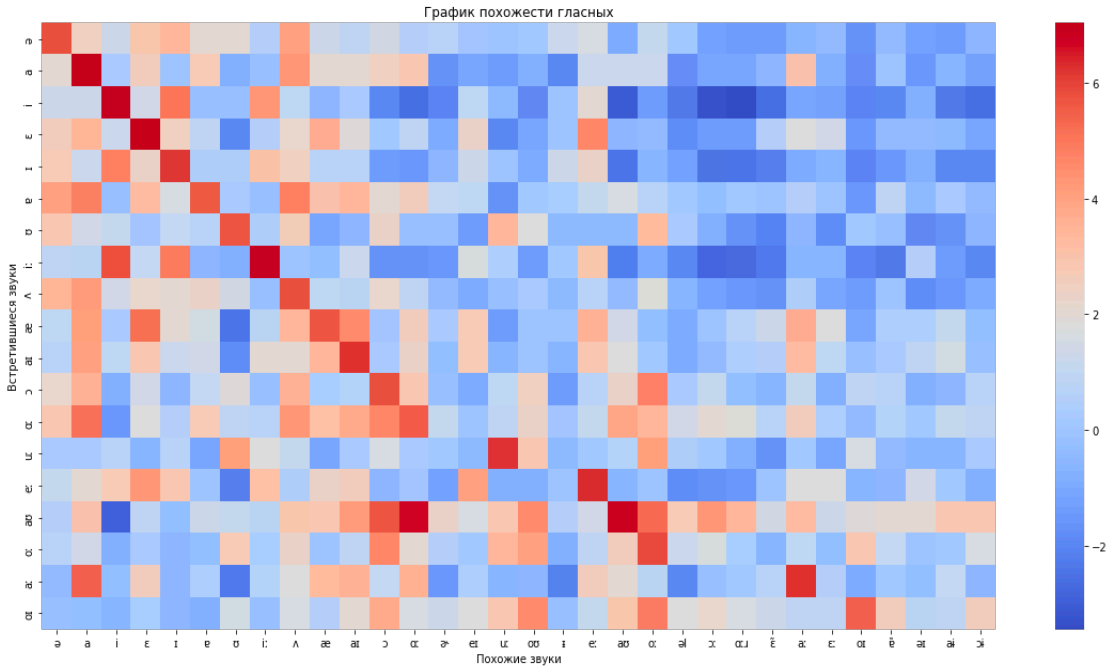
T. Nagamine, M. L. Seltzer, N. Mesgarani. Exploring how deep neural networks form phonemic categories // Proc. Interspeech. - P. 1912–1916. - 2015.

Yonatan Belinkov, James Glass. Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems // arXiv preprint arXiv:1709.04482. - 2017.

Приложение



Приложение 1. Тепловая карта сходства согласных. На оси Y отмечены согласные, попавшие в словарь модели wav2vec 2.0 на основе аудиозаписей русского языка. По оси X отмечены согласные, которые наиболее похожи на согласные из словаря. Чем краснее цвет квадрата, тем больше похожи два звука, на чьём пересечении лежит квадрат.



Приложение 2. Тепловая карта сходства гласных. На оси Y отмечены гласные, попавшие в словарь модели wav2vec 2.0 на основе аудиозаписей русского языка. По оси X отмечены гласные, которые наиболее похожи на гласные из словаря. Чем краснее цвет квадрата, тем больше похожи два звука, на чьём пересечении лежит квадрат.