# Использование обучения с подкреплением для решения задачи распознавания диктора в интерактивном режиме

Головин Вячеслав Сергеевич

2023

## Содержание

# Введение

Данная работа посвящена интерактивному подходу к решению задачи распознавания диктора. Оригинальный метод был предложен в [1], и значительная часть работы посвящена его описанию и практической реализации. С момента публикации этой статьи (2020 год) уже прошло достаточное количество времени, но она не стала популярной — по данным *Google Scholar* на момент написания этого отчёта она была процитирована 6 раз[1]. Тем не менее, нам (автору дипломной работы, его научному руководителю и коллегам из лаборатории Huawei CBG AI) она показалась заслуживающей внимания. На это есть ряд причин.

В первую очередь стоит отметить оригинальность предложенного подхода. Исторически большинство работ, в той или иной степени затрагивающие задачу распознавания диктора, посвящены способам как можно лучше определять диктора на основе уже существующих аудиозаписей. [здесь, наверное, нужно привести примеры таких работ] Рассматриваемая работа ставит проблему иначе — какие слова или фразы должен произнести диктор, чтобы уже существующая система смогла распознать его как можно быстрее и надёжнее. Чем-то такой подход напоминает концепцию активного обучения (*англ.* active learning) — разметки только тех данных, которые являются наиболее важными для решающей функции.

Другой причиной интереса к работе стала возможность её потенциального использования в конечном продукте — системе аутентификации пользователя на мобильном устройстве или персональном ассистенте. Предополагается, что такая система будет спрашивать пользователя произнести ту или иную фразу, пока она не станет уверена, что перед ней действительно находится настоящий владелец прибора. В таком случае логично делать не случайные запросы, а такие, которые позволят системе как можно быстрее идентифицировать пользователя.

Более подробное описание метода дано в главе 1. Следующая глава посвящена практической реализации описанного метода и полученным результатам. Глава 3 в свою очередь посвящена модификациям оригинального подхода, направленными на повышение точности и адаптации метода под сформулированную выше практическую задачу.

[1]Из этих цитат 1 приходится на кандидатскую диссертацию её первого автора.

# 1 Распознавание диктора в интерактивном режиме

## 1.1 Задача распознавания диктора

...

## 1.2 Интерактивный режим

...

# 2 Детали реализации и результаты

## 2.1 Данные для обучения и извлечение эмбеддингов

Здесь мы практически полностью повторяем описанный в [1] подход. Единственным (но очень существенным) отличием является использованная размерность эмбеддингов. Перед тем как перейти к обсуждению этого момента, расскажем про исходные данные.

Итак, для обучения и тестирования моделей мы использовали датасет TIMIT[2]. Он составлен из аудиозаписей речи 630 дикторов, говорящих на 8 основных диалектах американского английского языка. Эти дикторы поделены на обучающую (`train`) и тестовую (`test`) выборки, в первую входят 468 дикторов, во вторую — 162. Для обучения нейросетевых моделей мы также создавали валидационную выборку, в которую выделялись 20% дикторов из обучающей.

Каждый из дикторов произносит 10 фонетически насыщенных предложений. При этом 2 из 10 предложений являются общими для всех дикторов [2] , остальные 8 уникальны для каждого диктора. Такое разделение позволяет без особых затруднений подготовить данные, необходимые для описанной в 1.2 игры:

- 2 общих предложения можно использовать для получения аудиозаписей слов. Для этого разделим аудиозаписи этих предложений по временным отметкам, предоставленным создателями датасета. В результате получим 20 аудиозаписей слов[3] для каждого диктора.

[2]Общие предложения:
*She had your dark suit in greasy wash water all year.*
*Don't ask me to carry an oily rag like that.*

[3]Аналогично [1] мы не используем слово *an*.

- 8 уникальных для каждого диктора предложений можно использовать для получения голосовых подписей — эмбеддингов дикторов — просто при помощи усреднения эмбеддингов аудиозаписей этих предложений.

В качестве векторов признаков использовались эмбеддинги `x-vector` [3]. Весь процесс преобразования аудиозаписей в векторы признаков осуществлялся с помощью библиотеки Kaldi [4]. На первом этапе рассчитывались мел-частотные кепстральные коэффициенты[4] и производилось детектирование голосовой активности (*англ.* VAD — voice activity detection). Полученные векторы признаков поступали на вход предобученной нейронной сети [5]. В качестве эмбеддингов использовались данные со второго 512-мерного слоя.

Здесь, как уже было сказано ранее, мы отступаем от оригинальной работы [1], где использовались 128-мерные эмбеддинги. На это есть две причины. Во-первых, из приведенных в [1] комментариев неочевидно[5], как производилось понижение размерности. Во-вторых, мотивация такого преобразования тоже неочевидна. Уже первые проведенные нами эксперименты показали, что при использовании 512-мерных эмбеддингов точность идентификации оказывается существенно выше приведенных в [1] значений.

## 2.2 Обучение `Guesser`

...

## 2.3 Обучение `Enquirer`

...

## 2.4 Эвристическая модель выбора слов

...

[4]Параметры аналогичны использованным в [1] и определяются требованиями предобученной модели.

[5]Цитата: *We then process the MFCCs features through a pretrained X-Vector network to obtain a high quality voice embedding of fixed dimension 128, where the X-Vector network is trained on augmented Switchboard, Mixer 6 and NIST SREs.*

# 3 Модификации метода

## 3.1 От идентификации к верификации

`Enquirer` менять вообще не нужно, `Guesser` — совсем немного.

## 3.2 `CodebookEnquirer`

вроде работает

## 3.3 Добавление шума

обучается норм, результаты такие же

## 3.4 Альтернативные эмбеддинги

внезапно эмбеддинги 2017 года оказались не очень

# Заключение

все работает, но хотелось бы большего

# Список литературы


[1] M. Seurin, F. Strub, P. Preux и O. Pietquin, *A machine of few words – interactive speaker recognition with reinforcement learning*, 2020. arXiv: 2008.03127 [eess.AS].

[2] Garofolo, John S. и др., *TIMIT acoustic-phonetic continuous speech corpus*, 1993. DOI: 10.35111/17GK-BN40. url: https://catalog.ldc.upenn.edu/LDC93S1.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey и S. Khudanpur, «X-Vectors: Robust DNN embeddings for speaker recognition,» в *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, апр. 2018. DOI: 10.1109/icassp.2018.8461375. url: https://www.danielpovey.com/files/2018_icassp_xvectors.pdf.

[4] D. Povey и др., «The kaldi speech recognition toolkit,» в *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Catalog No.: CFP11SRW-USB, Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, дек. 2011.

[5] «SRE16 Xvector Model.» (2017), url: http://kaldi-asr.org/models/m3 (дата обр. 18.05.2023).