РАЗРАБОТКА СОВРЕМЕННОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ РУССКОЯЗЫЧНОЙ РЕЧИ ДЛЯ ДОМЕНОВ ТЕЛЕФОНИИ И YOUTUBE

Обухов Д.С.[1]

(*Новосибирский государственный технический университет, Новосибирск; Dasha.AI, Новосибирск*)

Описывается система, разработанная для распознавания русскоязычной речи. Мы фокусируемся на домене телефонных разговоров, когда на вход поступает одноканальный аудиосигнал с частотой дискретизации 8 кГц, полученный в условиях с повышенными шумами. Помимо основного домена телефонных разговоров, для обучения используются данные из видеохостинга YouTube. Рассматривается ряд акустических моделей, среди которых наиболее эффективной оказалась архитектура нейронной сети с временной задержкой и матричной факторизацией. Кроме того, приводятся результаты экспериментов по влиянию информации о спикере. Также в работе рассматривается применение различных техник аугментации и показывается, что применение таких техник аугментации, как реверберация, изменение скорости и громкости сигнала, маскирование частотных и временных характеристик, существенно повышает качество распознавания. На валидационном наборе данных телефонии достигнута ошибка обучения на словах WER 29.17.

Ключевые слова: распознавание речи, русскоязычная речь, акустическая модель, языковая модель, аугментация звука, эмбеддинги спикера.

1. Введение

Распознавание речи телефонных разговоров - задача, которая вызывает большой интерес в настоящий момент. Как и в общем случае для распознавания речи в данном домене используются разные подходы: комплексные, в которых транскрипции получаются с помощью одной натренированной модели, и гибридные, в которые обучается несколько моделей, как правило, акустическая, языковая модели и модель произношения.

Исторически, гибридные подходы появились раньше. Сперва это были системы на основе скрытых марковских цепей, в которых для моделирования вероятностей наблюдений исполь-

[1] *Обухов Дмитрий Сергеевич, аспирант, инженер-исследователь* (*bstodin@gmail.com*).

зовались смеси гауссовских моделей (HMM-GMM системы) [3]. Затем глубокие нейронные сети (DNN) заменили GMM [10]. Вместе с тем, как нейронные сети стали развиваться для гибридных подходов [14,15], популярность стали набирать комплексные системы. Рост вычислительных мощностей и новые алгоритмы машинного обучения позволили полностью обновить парадигму традиционных подходов распознавания речи. Комплексные подходы разделяются на три основные группы: подходы на основе рекуррентных сетей (RNN) [2,8], подходы на основе CTC функции потерь [1,9], подходы на основе механизма внимания (Attention) [4,6]. Сейчас нет однозначного фаворита среди гибридных и комплексных подходов по распознаванию речи, и каждая из этих техник представляет интерес в научном сообществе и активно продвигается вперед.

В данной статье рассматривается гибридный подход для решения задачи распознавания телефонной русскоязычной речи. Вместе с тем, видеохостинг youtube - источник большого количества размеченных записей, поэтому для обучения и валидации используются данные из двух доменов - youtube и телефонные звонки.

Одной из целей данной работы было сравнение архитектуры нейронной сети с временной задержкой и матричной факторизацией (TDNN-F) [15], которая считается одной из лучших на сегодняшний день акустической моделью в гибридном подходе, с другими архитектурами на базе нейронных сетей с временной задержкой (TDNN).

Для распознавания речи можно и нужно использовать информацию о спикере. Используя технику, предложенную в работе [17], мы закодировали эту информацию в вектора фиксированной размерности - эмбеддинги спикера i-vectors, и исследовали влияние этой информации на качество распознавания. В телефонии много особенностей, в частности сигнал содержит много шумов и есть ограничение на частоту дискретизации. В работах [11,12,13] показано, что правильное применение аугментации сильно повышает качество распознавания зашумленной речи. Ав-

торами данной работы, показано, что техники аугментации, такие как изменение скорости, изменение громкости, наложение шумов и эхо-эффекта, маскирование частотных и временных характеристик могут быть успешно применены для распознавания русскоязычной речи.

Структура статьи следующая: во второй главе описываются акустические модели, которые рассматриваются в данной работе. В третьей главе приводится описание рассмотренных техник аугментации. В четвертой главе отображаются результаты экспериментов. Заключение подводится в пятой главе.

2. Акустическая модель

Акустическая модель играет является одним из ключевых компонентов в гибридном подходе распознавания речи. Роль акустической модели заключается в построении последовательности сенонов для заданной аудиозаписи. Сеноны - это промежуточные представления между символами алфавита и звуками, которое языковые модели трансформируют в текст. В данной работе были рассмотрены акустические модели со следующими архитектурами:

- TDNN-F;
- TDNN-LSTM;
- TDNN-CNN;
- TDNN-Attention.

TDNN-F архитектура [15] на сегодняшний день показывает лучшие результаты в англоязычном домене. Это модификация TDNN архитектуры [14], в которой удалось сократить число параметров, за счет того, что матрица параметров раскладывается в произведения двух матриц с меньшей внутренней размерностью, причем первая из них является полу-ортогональной, за счет чего не происходит потери информации при снижении размерности.

TDNN-LSTM архитектура [5] включает блоки TDNN слоев и слоев с долгой краткосрочной памятью (LSTM). Поскольку речь

от природы является динамичным процессом, кажется естественным использовать рекуррентные нейронные сети (RNN) для распознавания. LSTM слои позволяют включить преимущества RNN сетей.

TDNN-CNN архитектура [7] включает в себя на нижних уровнях сверточные слои, а на более высоких TDNN слои. Таким образом нижний блок из конволюционных слоев позволяет извлечь более высокоуровневые признаки из аудио сигнала.

TDNN-Attention включает в себя слои с механизмом самовнимания (self-attention) [18]. В распознавании речи применяется т.н. “ограниченный” механизм само-внимания, так как только некоторый контекст слева и справа учитывается на каждой итерации по времени. Архитектура похожа на архитектуру TDNN-LSTM, в которой слои само-внимания заменили рекуррентные слои [16].

В данной работе со всеми акустическими моделями была использована статистическая языковая модель обученная на триграммах.

3. Аугментация данных

В данной работе были применены следующие техники аугментации:

- реверберация;
- изменение скорости аудиосигнала;
- изменение громкости аудиосигнала;
- маскирование частотных и временных характеристик.

Под реверберацией понимался наложение эхо на аудиозапись. Для применения реверберации был использован набор данных RIRS NOISES [12]. Размеры комнаты, в которой стимулировался эхо эффект, равномерно распределены от 1 до 30 метров. Реверберация была применена к двум копиям исходного набора данных для обучения, тем самым увеличив набор данных для обучения вдвое.

Изменение скорости выполнялось тремя способами, как это делали [11], с коэффициентами 0.9, 1, 1.1, увеличивая набор данных обучения в три раза.

Изменение громкости происходило коэффициентов, равномерно распределенным в интервале от 0.5 до 2.

Техника маскирования частотных и временных характеристик была предложена в работе [13]. В отличии от других типов аугментации, маскирование частотных и временных характеристик применялось не к исходному аудио сигналу, а к мел спектрограмме.

4. Эксперименты данных

Все эксперименты проводились на машине со следующие конфигурацией: CPU: AMD Ryzen Threadripper 2950X 16-Core Processor; GPU: 4x NVidia GeForce RTX 2080.

Аудио данными для обучения и валидации являлись 8 кГЦ, 16-бит, одноканальные wav файлы.

Весь набор данных обучения включал примерно 800 часов записей из двух доменов - youtube и телефония, в соотношении примерно 75 к 25. Для валидации использовались примерно 4 часа записей из домена youtube и около 1 часа записей из домена телефонии.

Для экспериментов была использована лишь десятая часть этого корпуса, чтобы сэкономить на времени обучения.

4.1. СРАВНЕНИЕ АКУСТИЧЕСКИХ МОДЕЛЕЙ

Таблица 1. Сравнение архитектур акустических моделей

	WER, youtube	WER, телефония
TDNN-F	25.11	32.12
TDNN-LSTM	29.31	37.75
TDNN-CNN	25.08	31.96
TDNN-Attention	31.54	38.44

Для оценки качества распознавания речи традиционно используется метрика Word Error Rate (WER) [19]. В таблице 1 приведены результаты WER для рассмотренных в главе 2 акустических моделей. Валидация проходила на двух доменах - youtube и телефония. Наиболее успешными оказались результаты TDNN-F и TDNN-CNN моделей. Для дальнейших экспериментов была выбрана архитектура TDNN-F.

4.2. ВЛИЯНИЕ ДОПОЛНИТЕЛЬНОЙ ИНФОРМАЦИИ О СПИКЕРЕ

Следуя предложению работы [17], были исследованы i-vectors, которые позволяют учитывать информацию о спикере и канале. Для эксперимента с i-vectors была выбрана TDNN-F модель.

Таблица 2. Влияние дополнительной информации о спикере, извлеченной в виде i-vectors

	WER, youtube	WER, телефония
Без i-vectors	25.11	32.12
С использование i-vectors	25.07	31.51

Для дальнейших экспериментов было решено использовать модели с i-vectors.

4.3. ПРИМЕНЕНИЕ АУГМЕНТАЦИИ

Для эксперимента была использована TDNN-F акустическая модель и i-vectors. В таблице 3 приведены результаты с применением различных техник аугментации для тренировочного набора данных.

Применение аугментаций наложения реверберации, изменения скорости и изменения громкости по отдельности не приносит существенного улучшения, поэтому базовый вариант, включает в себя сразу три техники аугментации.

Как видно из результатов, применение аугментации сильно влияет на качество распознавания. Наилучший эффект достигается, при применении всех видов аугментации - наложения ревер-

Таблица 3. Влияние аугментации на качеств распознавания. rvb - реверберация, sp - изменение скорости, vp - изменение громкости, spec - наложение частотных и временных масок

	WER, youtube	WER, телефония
Без аугментации	27.78	37.18
rvb + sp + vp	25.07	31.51
rvb + sp + vp + spec	24.17	29.93
rvb (x2) + sp + vp + spec	24.24	29.4
rvb (x3) + sp + vp + spec	24.11	30.26

берации, изменение скорости, громкости и маскирования частотных и временных характеристик. При этом дальнейшее увеличение обучающего набора данных, путем дублирования аугментации реверберации не оказывает значительного эффекта.

Наложение реверберации и изменения скорости увеличивает набор данных обучения, поэтому такие модели обучаются на порядок дольше. По этой причине для обучения модели на полном объеме данных была использована реверберация только на одну копию исходных данных.

Маскирования частотных и временных характеристик применяется на лету и не существенно влияет на время обучения.

4.4. ВРЕМЯ ОБУЧЕНИЯ

Как было отмечено выше, некоторые техники аугментации увеличивает объем обучающей выборки, что приводит к увеличению времени обучения системы. Поэтому практический интерес представляет и время, необходимое для обучения акустических моделей с применением различных техник аугментации. В таблице 4, приведено время на обучение основных акустических моделей, из экспериментов выше:

Стоит отметить, что использование i-vectors не существенно влияет на время обучения акустической модели, но при этом в данной таблице не учитывается время, необходимое на обучение модели, извлекаемой i-vectors, а также время на извлечение самих i-vectors. Тем не менее, это время на порядок меньше времени

Таблица 4. Время обучения акустических моделей

	Время обучения (секунды)
TDNN-LSTM, rvb + sp + vp	36 596
TDNN-CNN, rvb + sp + vp	36 437
TDNN-Attention, rvb + sp + vp	25 000
TDNN-F	5 772
+ rvb + sp + vp	40 383
+ i-vectors	40 708
+ spec	41 063

обучения акустической модели.

4.5. СРАВНЕНИЕ С ДРУГИМИ РЕШЕНИЯМИ

Сравнение проводилось на тех же наборах данных, что и эксперименты: youtube и телефония. В сравнении принимали участие модели распознавания речи от Яндекса и модель Н. Шмырева - VOSK[2] и рассмотренная в нашей работе TDNN-F модель с i-vectors и аугментацией в виде наложения реверберации, изменения скорости и громкости сигнала и маскирования частотных и временных характеристик.

Модель распознавания от Яндекса была использована с кодовым названием “general:rc”, последняя на момент проведенного исследования. Распознавание выполнялось в потоковом режиме.

Единственное отличие нашей финальной модели, от тех, которые были рассмотрены в экспериментах - все имеющиеся данные были использованы для обучения.

Таблица 5. Сравнение системы с другими решениями

	WER, youtube	WER, телефония
VOSK	42.88	52.22
Yandex STT	42.00	**28.22**
TDNN-F (our)	**22.63**	29.17

[2] *https://github.com/alphacep/vosk - VOSK Speech recognition toolkit*

5. Заключение

Была рассмотрена система распознавания русскоязычной телефонной речи в условиях с повышенными шумами. Было проведено исследование различных техник аугментации, таких как реверберация, изменение скорости и изменение громкости аудио сигнала, маскирование частотных и временных характеристик. Также было проведено сравнение различных архитектур акустических моделей, и исследовано влияние информации о спикере. Финальная система достигла ошибки обучения на словах WER 29.17. В будущем планируется поиск более продвинутых техник аугментации данных для интересующего домена, а также повышение качества данной модели за счет улучшений языковой модели.

Литература

1. AMODEI D., ANANTHANARAYANAN S., ANUBHAI R., BAI J., BATTENBERG E., CASE C., CASPER J., CATANZARO B., CHENG Q., CHEN G. *Deep speech 2: End-to-end speech recognition in english and mandarin* // International Conference on Machine Learning, 2016, pp. 173–182.
2. BATTENBERG E., CHEN J., CHILD R., COATES A., LI Y.G.Y., LIU H., ZHU Z. *Exploring neural transducers for end-to-end speech recognition* // IEEE Workshop on Automatic Speech Recognition and Understanding, 2017, pp. 206–213.
3. BOURLARD H.A., MORGAN N. *Connectionist Speech Recognition: A Hybrid Approach* // Kluwer Academic Publishers. Norwell, MA, USA, 1993.
4. CHAN W., JAITLY N., LE Q.V., VINYALS O. *Listen, attend and spell* // CoRR, vol. abs/1508.01211, 2015.
5. CHENG G., PEDDINTI V., POVEY D., MANOHAR V., KHUDANPUR S., YAN Y. *An exploration of dropout with lstms* // in Proceedings of Interspeech, 2017.
6. CHOROWSKI J., BAHDANAU D., SERDYUK D., CHO K., BENGIO Y. *Attention-based models for speech recognition* // in Annual Conf. on Neural Information Processing Systems, 2015, pp. 577–585.
7. GHAHREMANI P., MANOHAR V., POVEY D., KHUDANPUR S. *Acoustic modelling from the signal domain using cnns* // in To appear in Interspeech 2016. IEEE, 2016.
8. GRAVES A. *Sequence transduction with recurrent neural networks* // arXiv:1211.3711 (arXiv preprint), 2012.
9. GRAVES A., FERNANDEZ S., GOMEZ F., SCHIDHUBER J. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks* // Inter. Conf. on Machine Learning, 2006, pp. 369–376.

10. HINTON G., DENG L. *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups* // Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82 –97, Nov. 2012.
11. KO T., PEDDINTI V., POVEY D., KHUDANPUR S. *Audio augmentation for speech recognition* // in Proceedings of INTERSPEECH, 2015. [Online]. Available: http://www.danielpovey.com/files/2015 interspeech augmentation.pdf
12. KO T., PEDDINTI V., POVEY D., SELTZER M., KHUDANPUR S. *A study on data augmentation of reverberant speech for robust speech recognition* // in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5220–5224.
13. PARK D.S. *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition* // arXiv e-prints, 2019.
14. PEDDINTI V., POVEY D., KHUDANPUR S. *A time delay neural network architecture for efficient modeling of long temporal contexts* // Proceedings of INTERSPEECH, 2015.
15. POVEY D., CHENG G., WANG Y., LI K., XU H., YARMOHAMADI M., KHUDANPUR S. *Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks* // Proceedings of the 19th Annual Conference of the International Speech Communication Association, INTERSPEECH 2018, Hyderabad, India, sep 2018.
16. POVEY D., HADIAN H., GHAHREMANI P., LI K., KHUDANPUR S. *A time-restricted self-attention layer for asr* // in ICASSP, 2018.
17. SAON G., SOLTAU H., NAHAMOO D., PICHENY M. *Speaker adaptation of neural network acoustic models using i-vectors* // in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013, pp. 55–59.
18. VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A.N., KAISER L., POLOSUKHIN I.

Attention is all you need // In Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

19. WANG Y., ACERO A., CHELBA C. *Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy* // IEEE Workshop on Automatic Speech Recognition and Understanding. St. Thomas, 2003.

SPEECH RECOGNITION SYSTEM FOR RUSSIAN TELEPHONE AND YOUTUBE SPEECH

Dmitry Obukhov, Novosibirsk State Technical University, Novosibirsk; Dasha.AI, Novosibirsk, post-graduate student, ml researcher (bstodin@gmail.com).

Abstract: We describe a system designed to recognize Russian-language speech. Our focus is on the domain of telephone conversations, when a single-channel noisy audio signal with a sample rate of 8 kHz is received at the input. A number of acoustic models are considered, among which the time-delay neural network with factorization (TDNN-F) architecture turned out to be the most effective. In addition, we conduct experiments on the influence of speaker information. It is also shown that the use of augmentation techniques such as reverb, changing the speed and volume of a signal, masking frequency and time characteristics significantly increase the quality of recognition. We achieve word error rate 29.17 on our validation dataset.

Keywords: speech recognition, russian-language speech, acoustic model, language model, speech augmentation, speaker embeddings.

УДК 004.934.1
ББК 32.813

Статья представлена к публикации
членом редакционной коллегии ...

Поступила в редакцию ...
Дата опубликования ...