Fast adaptation of component automatic speech recognition systems

Daniil Grebenkin (d.grebenkin@g.nsu.ru)
Novosibirsk State University
Laboratory of Applied Digital Technologies of IMC MMF NSU
Novosibirsk, Russia

Abstract

The purpose of this work was to figure out advantages and disadvantages of different fast component ASR adaptation methods on language model level. This feature is extremely demanded by NLP and AI experts because it can make some general ASR models more specialized. This paper provides the analysis of creation and application efficiency of different adapted vosk-model-ru-0.22 versions to audiotracks with new lexicon from the position of the improvement of the quality of recognition and amounts of used computing resources.

Key words: speech recognition; language model; adaptation; N-gram smoothing; masked language modeling.

Быстрая адаптация компонентных систем распознавания речи

Д.В. Гребенкин (d.grebenkin@g.nsu.ru)
Новосибирский национальный исследовательский государственный университет
Лаборатория прикладных цифровых технологий ММЦ ММФ НГУ
Новосибирск, Россия

Аннотация

Целью данной работы было оценить преимущества и недостатки различных методов быстрой адаптации компонентных систем распознавания речи на уровне языковых моделей. Такая особенность является крайне востребованной специалистами в области обработки естественного языка или искусственного интеллекта, поскольку она позволяет сделать модели распознавания речи широкого профиля более специализированными. В статье анализируются создание и эффективность применения различных адаптированных модификаций модели распознавания речи vosk-model-ru-0.22 на аудиозаписях, содержащих лексику из новой предметной области, с точки зрения улучшения качества распознавания и количества используемых вычислительных ресурсов.

Ключевые слова: распознавание речи; языковая модель; адаптация; сглаживание N-грамм; masked language modeling.

1 Введение

Задача адаптации системы распознавания речи состоит в том, чтобы научить модель распознавать новые слова из какой-либо предметной области с учетом их фонетических и фонологических особенностей, для этого могут использоваться специальные выборки фонем для конкретных языков, например [1], различные методы представления акустических признаков из звукового сигнала и т.д.. Современные системы распознавания речи можно поделить на два типа в зависимости от используемого подхода в их структуре: «классический» компонентный (модульный) подход и так называемый «end-to-end» (интегральный или «сквозной») подход.

Компонентный подход подразумевает, что система распознавания речи состоит из нескольких модулей, которые обучаются независимо друг от друга и затем объединяются для получения результата. В качестве примеров такого подхода можно назвать системы Kaldi[2], CMU Sphinx[3].

Система на основе end-to-end подхода чаще всего состоит из модуля выделения признаков из звукового сигнала и нейронной сети, которая на выходе должна выдавать цепочку слов. Такая структура предполагает, что одна цельная модель (нейросеть) преобразует звуковые признаки в цепочку фонем и затем в цепочку слов. Примерами таких моделей могут быть wav2vec2[4], QuartzNet[5].

Структура компонентных систем распознавания речи, в отличие от end-to-end, предоставляет возможность модифицировать каждый модуль без пересоздания всей системы целиком — без переобучения нейросети в примере системы с end-to-end структурой, поэтому их адаптацию можно назвать «быстрой». Типы данных, принимаемые на вход акустической или языковой моделью при распознавании отличаются, что позволяет адаптировать систему к новым словам с учетом тех данных, которые имеются у исследователя: возможна адаптация на уровне языковой модели при наличии слова и его транскрипции, но отсутствия аудиозаписи с произнесением этого слова в речи; возможна адаптация на уровне только акустических моделей при наличии информации об акустических особенностях звукового сигнала. Такая структура имеет как свои преимущества, так и недостатки в сравнении с end-to-end подходом: обучение с использованием априорных статистических данных (N-грамм, различных соотношений букв и аллофонов) является методом регуляризации модели, и в то же время такие статистические ограничения могут не соответствовать реальной речевой ситуации (словарь произношений может не отражать некоторые речевые особенности, языковые модели могут представлять не тот дискурс); качество распознавания речи системой с end-to-end подходом зависит только от корпуса, на котором обучается модель, поэтому для обучения такой модели требуется гораздо больше данных и ее сложнее регуляризовать. Еще одним недостатком компонентного подхода является необходимость в понимании работы всей системы и каждого модуля в отдельности пользователем, что для неспециалиста стало бы дополнительным препятствием. Тем не менее, модульный подход позволяет опытным исследователям в области компьютерной лингвистики относительно быстро адаптировать систему к новым словам и их особенностям.

Несмотря на наличие возможности адаптации акустической модели, чаще всего отдается предпочтение именно работе с языковыми моделями и на это есть несколько причин:

- для обучения и последующей адаптации языковых моделей возможно формирование сколь угодно больших текстовых корпусов, что является относительно несложной задачей, поскольку в Сети имеется достаточное количество ресурсов с текстами в открытом доступе. Формирование же акустических корпусов, особенно таких, где звукозаписи сопровождаются разметкой (текстовыми аннотациями) — это гораздо более трудоемкая и ресурсоемкая задача;
- сферы применения распознавания речи гораздо сильнее отличаются в лексическом смысле — словарём и синтаксической структурой высказываний — чем в акустико-фонетическом смысле: в задаче адаптации модели к словам конкретного языка зачастую не так важны тембр или особенности артикуляции отдельных носителей этого языка, принимается тот факт, что у большинства носителей одинаковое устройство артикуляционного аппарата, способы произнесения звуков и т.д.

Данная работа сравнивает эффективность применения различных способов быстрой адаптации компонентной системы распознавания спонтанной речи к новой лексике на уровне языковых моделей, оценивает их с точки зрения улучшения качества распознавания и затраченных ресурсов.

2 Методы адаптации языковых моделей

Языковой моделью (LM, language model) называют распределение вероятностей над последовательностями слов. Выбор метода адаптации языковой модели зависит от ее типа: на основе N-грамм и на основе нейронных сетей различных архитектур.

2.1 Адаптация языковых моделей на основе N-грамм

N-граммы — последовательности, состоящие из N-слов. На практике N-граммная модель[7] содержит пары <N-грамма, вероятность ее появления>. Эти вероятности мы можем использовать для предсказания вероятности произвольной цепочки слов. Для добавления новых N-грамм обычно используют интерполяцию[8] новой и исходной моделей:

1. создается новая N-граммная языковая модель из корпусов текстов, содержащих слова из предметной области, к которой необходимо исходную модель адаптировать; при этом возможно использование различных техник сглаживания N-грамм [7];
2. N-граммы комбинируются с N-граммами из исходной модели, создается новая языковая модель — адаптированная версия исходной — с новыми вероятностями[9].

2.2 Адаптация нейросетевых языковых моделей

Современными нейросетевыми языковыми моделями являются нейронные сети прямого распространения, которые были впервые описаны в работе Йошуа Бенджио [10], рекуррентные нейронные сети [11], и трансформеры типа BERT [12] и GPT[13] и Transformer-XL[6]. Для адаптации нейросетей к новой предметной области обычно используется так называемый «fine-tuning»[14], или «дообучение»: «размораживаются» последние слои нейросети и обучаются на данных исследователя. Использование fine-tuning рекуррентных нейронных сетей для анализа тональности текста и текстовой классификации позволило уменьшить количество ошибок моделей на текстах из различных корпусов [15]. В работе [16] сравниваются методы извлечения признаков (feature extraction) и fine-tuning для адаптации языковых моделей BERT и ELMo [17] к разным новым задачам. Авторы отмечают, что достоинства этих подходов отличаются друг от друга: feature extraction имеет вычислительные преимущества, а fine-tuning лучше подходит для того чтобы скорректировать слои нейросети, которые отвечают за абстрактные представления модели для специальной предметной области. При анализе авторами было отмечено, что BERT показывает лучшие результаты при «дообучении» на новых данных. В рамках работы [18] была впервые создана модель BERT для русского языка — RuBERT[19] путем обучения универсальных версий BERT под данный домен, и подробно описан метод переноса знаний.

3 Адаптация модели vosk-model-ru-0.22

Для того, чтобы экспериментально сравнить качество различных модификаций компонентной системы распознавания речи, адаптированными под определенную предметную область, использовались модель распознавания речи vosk-model-ru-0.22 [20] и RuBERT (для рескоринга). Система распознавания речи vosk [21] является

удобной надстройкой для Kaldi, существует множество моделей распознавания речи для vosk для более чем 20 разных языков и диалектов.

В качестве предметной области для адаптации были выбраны искусственные нейронные сети. Текстами для адаптации послужили статьи из Википедии[22], содержащие термины, которые можно отнести к данной предметной области.

3.1 Адаптации N-граммной языковой модели

Для добавления новых N-грамм в языковую модель vosk-model-ru-0.22 было решено использовать метод из п. 2.1. При этом при подсчете N-грамм для новой модели были использованы разные типы и параметры сглаживания [23] для создания ненулевых значений вероятностей у новых встречающихся N-грамм. В результате были созданы разные версии 4-граммных адаптированных языковых моделей (таблица 1).

Модель	1	2	3	4
Тип сглаживания	Уиттена-Белла [24]	Лапласа(add-one) [7]	Лапласа(add-k) [7]	Кнезера-Нея [25]
Значение	——	1	0.1	——

Таблица 1: Различные версии адаптированных языковых моделей vosk-model-ru-0.22

Для создания разных версий новых моделей, их интерполяции с исходной и подсчета перплексии получившихся адаптированных моделей использовался инструмент SRILM [26]. После создания новой модели с помощью инструментов Kaldi обновлялся декодер HCLG-граф [27], использовались материалы работы [28].

3.2 Fine-tuning RuBERT

Модель RuBERT была «дообучена» на текстах тестовой выборки с помощью библиотеки transformers[29] для задачи Masked Language Modeling (MLM) [30]. Обновленный RuBERT использовался для рескоринга (N-best rescoring) в случае наличия нескольких наиболее вероятных вариантов предсказаний языковой модели для определенных временных отрезков, с поочередного маскирования каждого слова (токена) в цепочке слов. Пример для одной аудиозаписи:

1. при распознавании аудиозаписи для некоторых фрагментов времени языковая N-граммная модель предлагает несколько вариантов с различной степенью «уверенности» (параметр «confidence») в порядке уменьшения: {'confidence': 244.932404, 'text': ' кубань спасибо'}, {'confidence': 242.95311, 'text': ' кубарь спасибо'}
2. в каждом варианте поочередно маскируется каждое слово (токен), считается распределение вероятностей слов модели для данной позиции, извлекается вероятность слова, которое изначально было под маской и добавляется в список результатов для данной гипотезы, получившиеся вероятности перемножаются, образуя таким образом значение для ранжирования вариантов;
3. из списка вариантов выбирается вариант с наибольшим значением, который и является результатом распознавания.

4 Тестирование

Тестовой выборкой для оценки адаптированных моделей распознавания речи стали фрагменты аудиозаписей (формата моно, WAV PCM, с частотой дискретизации 16000 Гц) лекций и их текстовые расшифровки университетского курса «Методы и алгоритмы компьютерной лингвистики»[35].

В качестве метрики качества предсказания слов в последовательности использовался критерий перплексии [31]. Тестовая выборка состояла из 113 предложений, 1742 слов, результаты представлены в таблице 2.

Модель	N-граммная модель в составе vosk-model-ru-0.22	1	2	3	4
Перплексия (PP)	1256.9	907.2	973.3	942.6	918.7

Таблица 2: Значения перплексии исходной языковой модели и её адаптированных версий

Модификации с RuBERT создавались на основе модели 1, как модели с наименьшим значением перплексии на тестовой выборке. Для оценки качества распознавания использовались показатели Word Error Rate (WER) и Character Error Rate(CER) [32] (таблица 3). Тестовая выборка состояла из нескольких фрагментов первой лекции курса [35] общей продолжительностью 15 минут. Аудиофайлы (моно, частота дискретизации 16000 Гц, формат WAV PCM) были выложены в отдельный репозиторий на портале GitHub[36].

Модель	vosk-model-ru-0.22	1	2	3	4	Модель 1 **без RNNLM,** с обновленным RuBERT для рескоринга	Модель 1 с обновленным RuBERT для рескоринга
Word Error Rate	0.216	0.212	0.213	0.212	0.209	0.250	0.216
Character Error Rate	0.094	0.087	0.089	0.088	0.087	0.098	0.087

Таблица 3: Значения WER и CER модели распознавания речи vosk-model-ru-0.22 и ее адаптированных версий

Для того, чтобы понять, является ли полученное качество распознавания речи оптимальным для задачи адаптации относительно других подходов, было решено адаптировать end-to-end систему с помощью использования адаптированных версий языковых моделей (табл. 1) для рескоринга и оценить качество распознавания на той же тестовой выборке. В качестве такой модели использовалась wav2vec2-large-ru-golos[37], которая была обучена авторами на обучающей части речевого корпуса Sberdevices Golos [38] в течение 12 эпох алгоритмом Adam. Результаты тестирования приведены в таблице 4.

Модель	wav2vec2-large-ru-golos	wav2vec2-large-ru-golos +модель 1	wav2vec2-large-ru-golos +модель 2	wav2vec2-large-ru-golos +модель 3	wav2vec2-large-ru-golos +модель 4
Word Error Rate	0.592	0.604	0.607	0.606	0.600
Character Error Rate	0.178	0.211	0.211	0.210	0.210

Таблица 4: Значения WER и CER исходной модели распознавания речи wav2vec2-large-ru-golos и ее адаптированных версий

5 Выводы

В работе были проанализированы существующие типы систем распознавания речи, было определено, что для задачи адаптации наиболее универсальными системами являются компонентные, т.к. возможна отдельная адаптация разных модулей в зависимости от имеющихся видов данных — аудиозаписей или текстов, содержащих лексику из новой предметной области, при этом для адаптации требуются меньшие объемы данных, чем для моделей с end-to-end структурой вследствие наличия регуляризирующих статистических ограничений. Наиболее важным уровнем для адаптации был выбран уровень языковых моделей, т.к.: 1) формирование достаточно больших текстовых корпусов является более ресурсоемкой задачей; 2) сферы применения распознавания речи сильнее отличаются в лексическом смысле, чем в акустико-фонетическом.

Применение современных методов адаптации компонентных систем распознавания речи на уровне языковых моделей зависит от типа языковых моделей: интерполяция различных N-граммных или fine-tuning нейросетевых моделей (BERT, GPT-2, RNN). Экспериментально было показано, что оба метода применимы при использовании языковых моделей для формирования наиболее вероятных вариантов распознанного фрагмента речи и решения задачи нахождения лучшего варианта из них (N-best rescoring): различные модификации интерполированной модели с разными типами сглаживания улучшают ее способность к предсказанию следующего слова в цепочке слов тестовой выборки (уменьшение значения перплексии) и повышают качество распознавания слов из новой предметной области без потери качества (уменьшение значений Word Error Rate и Character Error Rate), в отличие от аналогичных модификаций на основе end-to-end модели; решение задачи N-best rescoring с помощью «дообучения» нейросетевых моделей типа BERT также улучшает качество распознавания речи, однако использование BERT без встроенной в vosk-model-ru-0.22 RNNLM показало более низкий результат, хотя слова из новой предметной области были распознаны. В дальнейшем планируется реализовать fine-tuning BERT, используя тренировочную выборку, которая содержит маркированные ошибочные результаты алгоритма распознавания речи, возможно использование моделей типа ruT5[33] для исправления грамматических ошибок распознавания.

С точки зрения количества затраченных ресурсов для процесса адаптации наименее ресурсоемкими являются нейросетевые модели, т.к. загрузка и интерполяция больших N-граммных моделей требует наличия большого объема оперативной памяти. При ограниченных вычислительных ресурсах компьютера для обучения нейросетевой модели и использования ее для рескоринга можно эффективно использовать параллельные вычисления на GPU при наличии технологии CUDA [34].

Библиография

[1] Yakovenko, O. S., Bondarenko, I. Y., Borovikova, M. N., & Vodolazsky, D. I. (2018). Algorithms for accentuation and phonemic transcription of Russian texts in speech recognition systems. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2018-May(17), 762-774.

[2] The Kaldi Speech Recognition Toolkit / Daniel Povey, Arnab Ghoshal, Gilles Boulianne et al. // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. — Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society, 2011. — Dec. — IEEE Catalog No.: CFP11SRWUSB.

[3] A. Dhankar, "Study of deep learning and CMU sphinx in automatic speech recognition," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 2296-2301, doi: 10.1109/ICACCI.2017.8126189.

[4] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2006.11477.

[5] Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., & Zhang, Y. (2019). QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1910.10261.

[6] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1901.02860.

[7] Jurafsky, D. & Martin, J. (2020). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. 3rd Edition draft.

[8] Liu, X., Gales, M. J. F., & Woodland, P. C. (2013). Use of contexts in language model interpolation and adaptation. In Computer Speech & Language (Vol. 27, Issue 1, pp. 301–321). Elsevier BV. https://doi.org/10.1016/j.csl.2012.06.004.

[9] Liu, Xunying & Gales, M.J.F. & Woodland, Philip. (2009). Use of contexts in language model interpolation and adaptation.. 360-363.

[10] Bengio, Y. & Ducharme, Réjean & Vincent, Pascal. (2000). A Neural Probabilistic Language Model. Journal of Machine Learning Research. 3. 932-938. 10.1162/153244303322533223.

[11] Mikolov, Tomas & Karafiát, Martin & Burget, Lukas & Cernocký, Jan & Khudanpur, Sanjeev. (2010). Recurrent neural network based language model. Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. 2. 1045-1048.

[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1810.04805.

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[14] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In Lecture Notes in Computer Science (pp. 194–206). Springer International Publishing. https://doi.org/10.1007/978-3-030-32381-3_16.

[15] Dai, A. M., & Le, Q. V. (2015). Semi-supervised Sequence Learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1511.01432.

[16] Peters, M. E., Ruder, S., & Smith, N. A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1903.05987.

[17] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1802.05365.

[18] Специализация языковых моделей для применения к задачам обработки естественного языка: специальность 05.13.17 "Теоретические основы информатики" : автореферат диссертации на соискание ученой степени кандидата физико-математических наук / Куратов Юрий Михайлович; [Московский физико-технический институт (национальный исследовательский университет)]. - Москва, 2020. - 25 с. : ил.. - Библиогр.: с. 19-25. - 16 экз.

[19] Kuratov, Y., & Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1905.07213.

[20] vosk-model-ru-0.22. URL: https://alphacephei.com/vosk/models/vosk-model-ru-0.22.zip (дата обращения: 11.05.2022).

[21] VOSK Offline Speech Recognition API. URL: https://alphacephei.com/vosk/ (дата обращения: 11.05.2022).

[22] Wikipedia - free encyclopedia. URL: https://ru.wikipedia.org (дата обращения: 11.05.2022).

[23] ngram-count. URL: http://www.cs.cmu.edu/afs/cs/project/cmt-55/lti/Courses/731/homework/HW8/srilm/man/html/ngram-count.1.html (дата обращения: 11.05.2022)..

[24] Witten, I. H. and T. C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory, 37(4):1085–1094.

[25] Goodman, J. (2001). A Bit of Progress in Language Modeling (Version 1). arXiv. https://doi.org/10.48550/ARXIV.CS/0108005.

[26] Working with n-grams in SRILM. Linguistics 165, Professor Roger Levy. 13 February 2015. URL: https://pages.ucsd.edu/~rlevy/teaching/2015winter/lign165/lectures/lecture13/lecture13_ngrams_with_SRILM.pdf (дата обращения: 11.05.2022).

[27] Decoding graph construction in Kaldi. URL: https://kaldi-asr.org/doc/graph.html (дата обращения: 11.05.2022).

[28] va-stepanov/vosk-model-ru-adaptation. URL: https://github.com/va-stepanov/vosk-model-ru-adaptation (дата обращения: 11.05.2022).

[29] transformers · PyPI. URL: https://pypi.org/project/transformers/ (дата обращения: 11.05.2022)..

[30] Alfaro, Felipe & Costa-jussa, Marta & Fonollosa, José. (2019). BERT Masked Language Modeling for Co-reference Resolution. 76-81. 10.18653/v1/W19-3811.

[31] Chen, Stanley & Beeferman, Douglas & Rosenfeld, Ronald. (2001). Evaluation Metrics For Language Models.

[32] Карпов Алексей Анатольевич, Кипяткова Ирина Сергеевна Методология оценивания работы систем автоматического распознавания речи // Приборостроение. 2012. №11. URL: https://cyberleninka.ru/article/n/metodologiya-otsenivaniya-raboty-sistem-avtomaticheskogo-raspoznavaniya-rechi (дата обращения: 11.05.2022).

[33] sberbank-ai/ruT5-base · Hugging Face. URL: https://huggingface.co/sberbank-ai/ruT5-large/tree/main (дата обращения: 11.05.2022).

[34] CUDA Toolkit - Free Tools and Training | NVIDIA Developer. URL: https://developer.nvidia.com/cuda-toolkit (дата обращения: 11.05.2022).

[35] Методы и алгоритмы компьютерной лингвистики. URL: https://vk.com/nsu_nlp (дата обращения: 11.05.2022).

[36] dangrebenkin/Computer_linguistics_test_dataset: Тестовая выборка аудиозаписей, состоящая из фрагментов лекции 1 курса «Методы и алгоритмы компьютерной лингвистики». URL: https://github.com/dangrebenkin/Computer_linguistics_test_dataset.git (дата обращения: 21.06.2022).

[37] bond005/wav2vec2-large-ru-golos · Hugging Face. URL: https://huggingface.co/bond005/wav2vec2-large-ru-golos (дата обращения: 21.06.2022).

[38] Nikolay Karpov, Alexander Denisenko, Fedor Minkin. Golos: Russian Dataset for Speech Research (2021), in Proceedings of Interspeech 2021, pages 1419-1423