УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2021 Том 21 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2021 Vol. 21 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2021-21-5-791-794
УДК 004.522

Архитектура системы полнотекстового поиска по речевым данным на основе глобального индекса

Олег Евгеньевич Петров

ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация
petrov-o@speechpro.com✉, https://orcid.org/0000-0001-8258-3171

Аннотация
Предложена архитектура системы полнотекстового поиска по речевым данным, основанная на глобальном индексе поиска, который объединяет в себе информацию обо всех фонограммах архива. Архитектура включает в себя два независимых блока: блок индексирования и блок формирования и выполнения поискового запроса. Обработка фонограмм осуществляется с помощью системы автоматического распознавания речи, которая использует лингвистический декодер на основе взвешенных преобразователей конечных состояний (WFST) для создания словных сетей. Последовательное формирование на основе данных блоков сетей спутывания и обратных индексов позволяет учитывать все словные гипотезы, сформированные в процессе декодирования. Предложенное решение расширяет границы применимости систем речевой аналитики на те случаи, когда пословная ошибка распознавания речи является высокой, например, при обработке фонограмм, полученных в сложных акустических условиях или на малоресурсных языках.

Ключевые слова
полнотекстовый поиск, речевая аналитика, поиск ключевых слов, поисковый индекс, автоматическое распознавание речи

Ссылка для цитирования: Петров О.Е. Архитектура системы полнотекстового поиска по речевым данным на основе глобального индекса // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 5. С. 791–794. doi: 10.17586/2226-1494-2021-21-5-791-794

The architecture of a system for full-text search by speech data based on a global search index

Oleg E. Petrov

STC-innovations Ltd., Saint Petersburg, 194044, Russian Federation
petrov-o@speechpro.com✉, https://orcid.org/0000-0001-8258-3171

Abstract
This paper presents the architecture of a system for full-text search by speech data based on a global search index that combines information about all speech recordings in the archive. The architecture includes two independent blocks: an indexing block, and a block for building and performing a search query. In order to process speech recordings, it uses an automatic speech recognition system (ASR) with a linguistic decoder based on weighted finite-state transducers framework (WFST), which generates word lattices. Lattices are sequentially converted to confusion networks and inverse indexes. It allows taking into account all the word hypotheses generated during decoding. The proposed solution expands the applicability of speech analytics systems for those cases when the word error rate is high, such as the processing of speech recordings collected under difficult acoustic conditions or in low-resource languages.

Keywords
full-text search, speech analytics, spoken term detection, search index, automatic speech recognition

For citation: Petrov O.E. The architecture of a system for full-text search by speech data based on a global search index. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 5, pp. 791–794 (in Russian). doi: 10.17586/2226-1494-2021-21-5-791-794

В колл-центрах и офисах продаж крупных компаний работают тысячи сотрудников, которые обрабатывают сотни тысяч обращений каждый день. В целях контроля качества обслуживания все разговоры записываются, формируя большие объемы аудиоданных, которые хранятся в постоянно пополняющихся архивах — специальных базах медиаданных. Речевая аналитика фонограмм обращения клиентов позволяет извлекать важную для бизнеса информацию: отзывы о предоставляемых товарах и услугах или оценки качества проведенных рекламных акций.

Современные системы автоматического распознавания речи дают возможность получать текстовое представление диалога клиента и оператора. Контроль качества обслуживания может сводиться к формированию каскада регулярно выполняемых поисковых запросов к базе аудиоданных.

Один из способов организации поискового индекса по текстам диалогов — использование обратного индекса слов [1]. Применительно к задачам речевой аналитики такой подход создает решения, обеспечивающие одновременно и высокую скорость индексации, и быстрый гибкий поиск, использующий специальный формат поисковых запросов. Точность поиска при этом ограничивается точностью выбранной системы автоматического распознавания речи. Применение текстового представления диалогов остается невозможным в тех случаях, когда пословная ошибка распознавания речевых данных остается высокой. К таким случаям можно отнести распознавание речи на малоресурсных языках или распознавание в сложных акустических условиях.

Поиск речевой информации по сетевому индексу фонограммы, содержащему все гипотезы, полученные лингвистическим декодером, повышает полноту результата за счет возможности обнаружения слов, распознанных с низким уровнем достоверности и не попавших в текстовый результат распознавания, но сохранившихся в словной сети декодера [2]. В современных системах используются декодеры на основе взвешенных преобразователей конечных состояний (Weighted Finite-State Transducers, WFST) в связке с акустическими моделями, построенными на различных архитектурах нейронных сетей [3, 4].

При переходе к работе с постоянно увеличивающимися объемами речевых данных, необходимо применять методы объединения сетевых индексов отдельных фонограмм в некоторые глобальные структуры — поисковые индексы, обеспечивающие возможность сублинейного времени поиска по речевым данным.

Словная сеть, полученная в результате работы WFST-декодера, представляет собой акцептор со словами на переходах и весами, полученными из акустической и языковой моделей. Словные сети не содержат циклов, имеют одно начальное состояние и одно финальное. При этом количество переходов из каждого состояния может быть произвольным и зависеть от настроек WFST-декодера и ограничений луча поиска [2]. Такая структура позволяет быстро обходить сеть, но в силу нерегулярной структуры плохо поддается индексации.

Вместо словной сети предполагается использовать более компактное представление порожденных гипотез — сети спутывания [5]. Сети спутывания обладают дополнительным свойством: все пути в сети спутывания от начального состояния к финальному проходят через все состояния. Дополнительно вводится ε-слово, означающее пустой переход, обозначающий пропуск слова или паузу. Набор переходов между парой соседних состояний называется бином. Каждый бин имеет временные метки начала и конца, которые размещаются в соответствующих состояниях.

Архитектура системы полнотекстового поиска по речевым данным представлена на рисунке и включает в себя два основных блока (выделены пунктирными линиями), связанных с использованием системы:

1) блок обработки фонограмм и построение индекса поиска — индексирование;

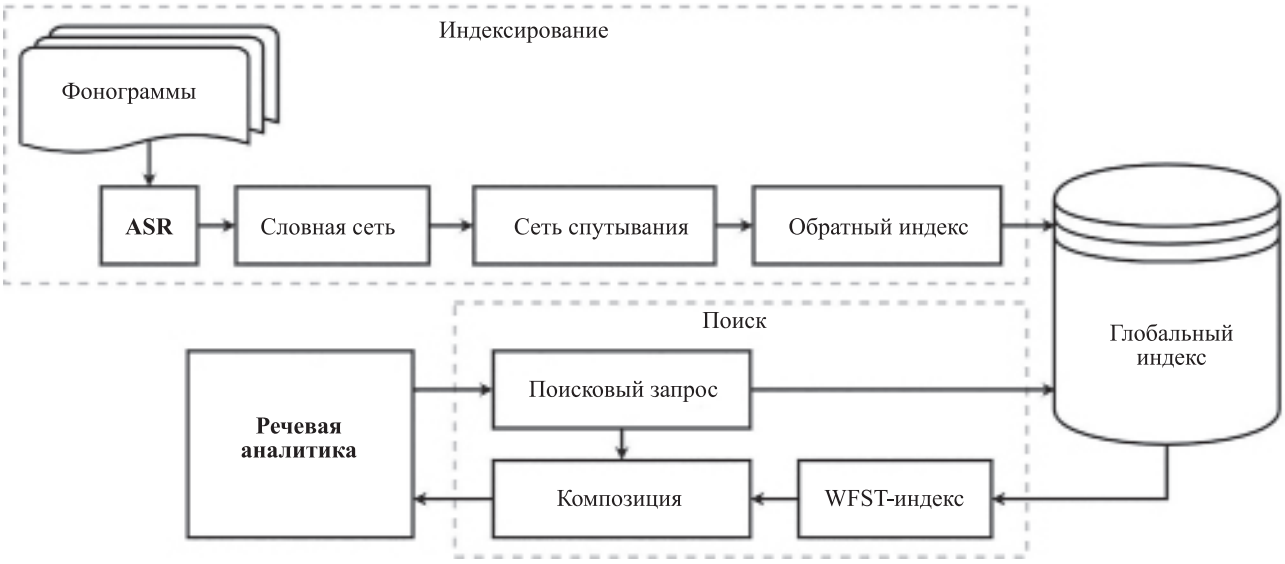

Рисунок. Архитектура системы полнотекстового поиска по речевым данным

Figure. The architecture of a system for full-text search by speech data

2) блок формирования и выполнения поискового запроса по сформированному индексу.

Процесс индексирования фонограмм может быть представлен следующим набором шагов.

Шаг 1. С помощью системы автоматического распознавания речи (ASR) на основе WFST-декодера формируется словная сеть, на основе которой строится сеть спутывания.

Шаг 2. Для сети спутывания строится обратный индекс. Обратный индекс сети спутывания — структура данных, реализующая интерфейс ассоциативного массива, в котором ключом является слово, в соответствие которому ставится список кортежей (n, p) для всех вхождений слова в сеть спутывания, где n — номер бина, в котором находится слово, а p — вероятность соответствующего перехода. Специальное ε-слово индексируется как обычное.

Шаг 3. Обратный индекс добавляется в глобальный индекс. Глобальный индекс поиска включает в себя информацию обо всех обработанных фонограммах и предоставляет механизмы для осуществления полнотекстового поиска по содержимому фонограмм. Считается обобщением обратного индекса сети спутывания. Каждый кортеж вхождения слова дополняется идентификатором фонограммы d: (d, n, p). Для быстрого обращения к словам в глобальном индексе используется дополнительный индекс на основе B-дерева [6].

Индексация фонограмм внутри глобального индекса осуществляется с помощью целочисленного идентификатора, по которому в отдельной структуре с прямой адресацией хранится метаинформация. Глобальный индекс делится на независимые разделы по N фонограмм, что позволяет параллельно обрабатывать группы фонограмм.

Процесс поиска фонограммы может быть представлен следующим набором шагов.

Шаг 1. На основе списка слов и словосочетаний, входящих в состав поискового запроса, строится словный акцептор Q так, чтобы все пути акцептора совпадали с искомыми фразами. Для получившегося акцептора Q дополнительно последовательно выполняются операции детерминизации и минимизации.

Шаг 2. Для ε-слова и каждого слова, входящего в запрос, из глобального индекса независимо для каждого раздела получаются обратные индексы.

Шаг 3. По полученным обратным индексам для каждого раздела строится WFST-индекс U, аналогично подходу, предложенному в [7], но с использованием только тех слов, для которых получены обратные индексы. Структура WFST-индекса, объединяющего несколько сетей спутывания, представляет собой трансдьюсер со словами в качестве входных меток и номерами документов и бинов в качестве выходных. WFST-индекс U строится таким образом, чтобы все пути по входным символам $\sum^*$ соответствовали всем возможным словосочетаниям, которые могут быть найдены в индексе. Ключевой особенностью является механизм построения трансдьюсера U. Благодаря тому, что WFST-индекс строится по отдельному разделу глобального индекса, в котором индексы фонограмм идут последовательно и ограничены фиксированным размером раздела, а количество бинов ограничено, можно использовать прямую адресацию для состояний. Для этого используется дополнительная память $O(B \cdot N)$, где B — максимальный индекс бина, а N — размер раздела глобального индекса.

Шаг 4. Выполняется композиция $R = Q \circ U$ для каждого раздела глобального индекса [8]. Входные символы трансдьюсера R — искомые слова, а выходные — номера документов и бинов конкретных слов. Все пути в трансдьюсере R соответствуют найденным словам и словосочетаниям из поискового запроса.

Предложенная архитектура системы полнотекстового поиска по речевым данным, включающая два набора компонентов для индексации и поиска, позволяет рассматривать каждый узел в отдельности. В качестве системы автоматического распознавания речи может быть применена любая система, использующая WFST-декодер, порождающий словные сети. Единожды обработанные фонограммы в виде обратных индексов размещаются в глобальном индексе, который хранит все гипотезы сети спутывания. За счет увеличения полноты результата, предложенная структура данных позволяет расширить границы применимости систем речевой аналитики для использования на малоресурсных языках или в сложных акустических условиях, где пословная ошибка распознавания остается высокой.

Литература

1. Zobel J., Moffat A. Inverted files for text search engines // ACM Computing Surveys. 2006. V. 38. N 2. P. 6–es. https://doi.org/10.1145/1132956.1132959
2. Saon G., Povey D., Zweig G. Anatomy of an extremely fast LVCSR decoder // Proc. 9th European Conference on Speech Communication and Technology. 2005. P. 549–552. https://doi.org/10.21437/Interspeech.2005-338
3. Mohri M., Pereira F., Riley M. Weighted finite-state transducers in speech recognition // Computer Speech and Language. 2002. V. 16. N 1. P. 69–88. https://doi.org/10.1006/csla.2001.0184
4. Laptev A., Andrusenko A., Podluzhny I., Mitrofanov A., Medennikov I., Matveev Y. Dynamic acoustic unit augmentation with BPE-dropout for low-resource end-to-end speech recognition // Sensors. 2021. V. 21. N 9. P. 3063. https://doi.org/10.3390/s21093063
5. Mangu L., Brill E., Stolcke A. Finding consensus in speech recognition: word error minimization and other applications of

References

1. Zobel J., Moffat A. Inverted files for text search engines. *ACM Computing Surveys*, 2006, vol. 38, no. 2, pp. 6–es. https://doi.org/10.1145/1132956.1132959
2. Saon G., Povey D., Zweig G. Anatomy of an extremely fast LVCSR decoder. *Proc. 9th European Conference on Speech Communication and Technology*, 2005, pp. 549–552. https://doi.org/10.21437/Interspeech.2005-338
3. Mohri M., Pereira F., Riley M. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 2002, vol. 16, no. 1, pp. 69–88. https://doi.org/10.1006/csla.2001.0184
4. Laptev A., Andrusenko A., Podluzhny I., Mitrofanov A., Medennikov I., Matveev Y. Dynamic acoustic unit augmentation with BPE-dropout for low-resource end-to-end speech recognition. *Sensors*, 2021, vol. 21, no. 9, pp. 3063. https://doi.org/10.3390/s21093063

confusion networks // Computer Speech and Language. 2000. V. 14. N 4. P. 373–400. https://doi.org/10.1006/csla.2000.0152
6. Lagogiannis G. Query-optimal partially persistent B-trees with constant worst-case update time // International Journal of Foundations of Computer Science. 2017. V. 28. N 2. P. 141–169. https://doi.org/10.1142/S0129054117500101
7. Mangu L., Kingsbury B., Soltau H., Kuo H.-K., Picheny M. Efficient spoken term detection using confusion networks // Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2014. P. 7844–7848. https://doi.org/10.1109/ICASSP.2014.6855127
8. Allauzen C., Riley M., Schalkwyk J. A filter-based algorithm for efficient composition of finite-state transducers // International Journal of Foundations of Computer Science. 2011. V. 22. N 8. P. 1781–1795. https://doi.org/10.1142/S0129054111009033

5. Mangu L., Brill E., Stolcke A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 2000, vol. 14, no. 4, pp. 373–400. https://doi.org/10.1006/csla.2000.0152
6. Lagogiannis G. Query-optimal partially persistent B-trees with constant worst-case update time. *International Journal of Foundations of Computer Science*, 2017, vol. 28, no. 2, pp. 141–169. https://doi.org/10.1142/S0129054117500101
7. Mangu L., Kingsbury B., Soltau H., Kuo H.-K., Picheny M. Efficient spoken term detection using confusion networks. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 7844–7848. https://doi.org/10.1109/ICASSP.2014.6855127
8. Allauzen C., Riley M., Schalkwyk J. A filter-based algorithm for efficient composition of finite-state transducers. *International Journal of Foundations of Computer Science*, 2011, vol. 22, no. 8, pp. 1781–1795. https://doi.org/10.1142/S0129054111009033

Автор

Петров Олег Евгеньевич — директор научно-исследовательского департамента, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация, SC 57211637406, https://orcid.org/0000-0001-8258-3171, petrov-o@speechpro.com

Author

Oleg E. Petrov — Director of the Research and Development Department, STC-innovations Ltd., Saint Petersburg, 194044, Russian Federation, SC 57211637406, https://orcid.org/0000-0001-8258-3171, petrov-o@speechpro.com

Статья поступила в редакцию 14.08.2021
Одобрена после рецензирования 21.08.2021
Принята к печати 17.09.2021

Received 14.08.2021
Approved after reviewing 21.08.2021
Accepted 17.09.2021