А.Г. ЯШИНА, Д.Е. ПРОЗОРОВ

Модель информационного поиска речевых документов по текстовому запросу на основе фонемного транскрибирования и TF-IDF меры

УДК 025.4.03

ФГБОУ ВПО «Вятский Государственный Гуманитарный Университет», г. Киров

ФГБОУ ВПО «Вятский Государственный Университет», г. Киров

В данной работе представлена модель информационного поиска речевых документов по текстовым запросам на основе фонемного транскрибирования распознанного текста и меры релевантности TF-IDF. Рассмотрен алгоритм автоматического фонемного транскрибирования, в котором используются вероятности появления фонем в зависимости от предыдущей фонемы и соответствующей буквы в слове. Приведены и проанализированы результаты поиска по значениям показателя R-точности для буквенного и фонемного представления распознанного содержания речевых документов на русском языке.

Введение

Речь является основной формой передачи информации для человека. Речевые документы представляют собой аудиофайлы, содержащие слитную спонтанную речь, например, радио-записи, записи лекций, конференций и бизнес встреч, аудио книги, архивы call-центров. Большинство информационных систем, работающих с мультимедийными документами, в том числе и речевыми, предоставляет возможность поиска на основе тэгов. Но данный подход требует ручного аннотирования, а также основывается на субъективном описании содержания документа. В связи с этим становится актуальной задача поиска речевых документов по содержанию.

Задача контекстного поиска речевых документов по текстовому запросу относится к области Spoken Document Retrieval (SDR) и

находится на стыке таких направлений как распознавание речи и информационный поиск. Основная проблема при решении данной задачи заключается в наличии большого количества ошибок распознавания, которые искажают содержание речевых документов, что отражается на эффективности поиска. Увеличение объема словаря системы распознавания речи позволяет снизить количество ошибок, но увеличивает время распознавания документов. Другим способом решения этой проблемы является применение методов поиска, учитывающих наличие ошибок распознавания речи в индексируемых документах [1].

Выделяют два основных подхода к реализации поиска речевых документов по текстовому запросу [1]. Первый подход основан на распознавании слитной речи в текст [2, 3]. Во втором подходе применяется фонемное распознавание речевых документов [4, 5]. Существуют методы, которые комбинируют оба подхода [6, 7]. Например, в [7] для определения слов используются решетки фонов, представляющие варианты фонетических транскрипций слова. Результаты экспериментов показывают [1], что методы, объединяющие оба подхода, превосходят методы, которые основаны на одном из подходов.

В данной работе представлена модель информационного поиска речевых документов по текстовым запросам на основе фонемного транскрибирования распознанного текста и меры релевантности TF-IDF.

Система контекстного поиска речевых документов

Основными этапами поиска речевых документов по текстовому запросу являются:

- распознавание содержания речевых документов,
- представление и индексирование содержания речевых документов,
- представление и индексирование текстового запроса в соответствии с формой представления содержания речевых документов,
- вычисление оценок релевантности документов запросу пользователя и вывод ранжированного списка документов.

Обобщенная схема системы контекстного поиска речевых документов приведена на рисунке 1.

Поиск выполняется в соответствии с некоторой моделью информационного поиска, которая подразумевает описание

- представления содержания документов,
- представления запросов пользователя,
- метода вычисления оценки релевантности документа запросу.

Существуют различные модели, которые изначально применялись в текстовом поиске. Наиболее известными являются булевая, вероятностная, векторная и языковая модели [8].

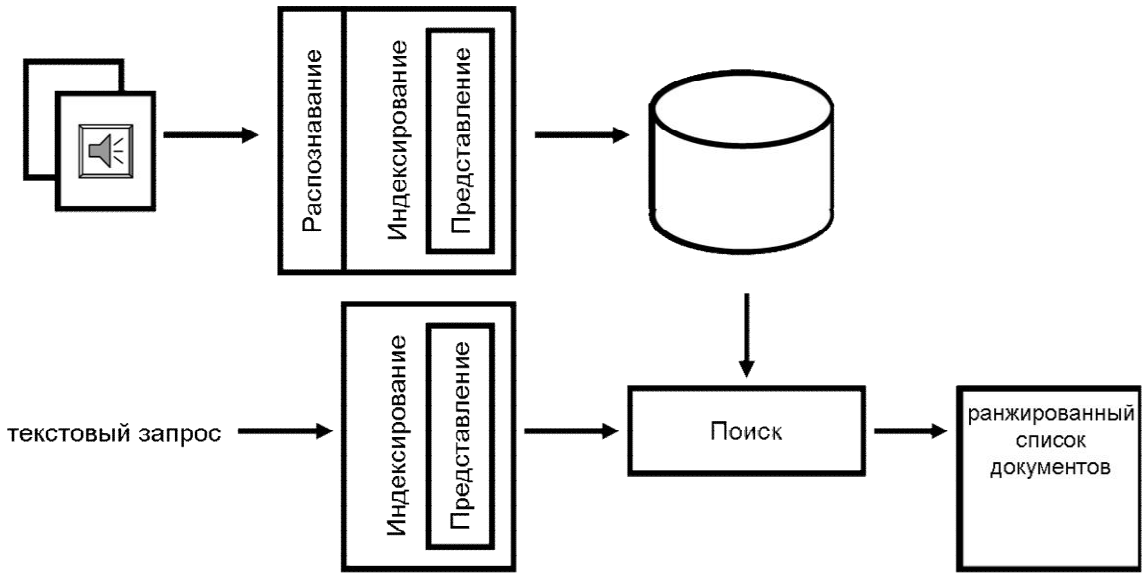

Рис 1. Общая схема системы контекстного поиска речевых документов по текстовому запросу.

Содержание речевого документа в системе контекстного поиска (рис.1) извлекается средствами распознавания слитной речи. При этом ошибки распознавания заключаются не только в замене, вставке или пропуске букв, но и неправильном определении пауз между словами. Так длинное слово может быть распознано как несколько коротких, что приводит к снижению эффективности поиска. Фонемное транскрибирование распознанного текста и запроса пользователя позволяет частично компенсировать указанные ошибки за счет учета близости фонем при сравнении транскрибированных слов [1,9].

После распознавания выполняется индексирование содержания речевого документа. Индексирование может включать изменение представления распознанного содержания речевого документа.

Например, в некоторых системах используются такие единицы представления как CVC-признаки [10], фонемы [4, 5, 11], *n*-граммы [12], что позволяет модифицировать способ вычисления оценки релевантности и повысить эффективность поиска. Проиндексированные документы сохраняются в базе данных системы (рис.1).

В процессе поиска текстовый запрос пользователя также индексируется и преобразуется к форме аналогичной форме представления распознанного содержания речевых документов.

Ранжирование документов подразумевает их упорядочивание по значению оценки релевантности запросу. Оценка релевантности основана на определении весов слов запроса для характеристики содержания документа. В системах текстового поиска для взвешивания слов часто применяется мера TF-IDF [8], которая неэффективна при поиске речевых документов по текстовому запросу из-за наличия ошибок распознавания. Поэтому предлагается её модификация.

Рассмотрим более подробно описываемую модель информационного поиска на основе фонемного представления содержания речевых документов и запросов, а также модифицированного алгоритма вычисления меры TF-IDF.

Постановка задачи

Содержание речевого документа d_k коллекции D, соответствует набору независимых распознанных слов $\left\{w_i^{d_k}\right\}$, $i=\overline{1,\,N_{d_k}}$, где N_{d_k} – количество слов документа d_k. Текстовый запрос Q представляет слова $\left\{q_j\right\}$, введенные пользователем, $j=\overline{1,\,N_Q}$, где N_Q – количество слов запроса. Необходимо вычислить значение релевантности каждого документа $d_k \in D$ запросу Q

$$r_k = F\left(d_k, Q\right), \tag{1}$$

где F – функция релевантности $d_k \in D$ запросу Q. Результатом поиска является список документов, ранжированный по r_k.

Содержание речевого документа рассматривается как объединение распознанных слов в одну фразу W_{d_k} без пробелов для исключения ошибок распознавания пауз между словами. Вариативность окончаний в словах сильно влияет на определение вхождения

слова запроса в содержание документа. В связи с этим при поиске используются основы слов, которые выделяются на этапе индексирования посредством стороннего алгоритма [13].

Метод фонемного транскрибирования

В описываемой модели информационного поиска используется фонемное представление распознанного содержания речевых документов и текстового запроса пользователя. Распознанный текст автоматически транскрибируется в последовательность фонем, для передачи особенностей произношения слов. Запросы пользователя транскрибируются аналогичным образом.

Сравнение слов запроса и содержания речевого документа в фонемном представлении позволяет частично компенсировать ошибки распознавания речи при ранжировании документов.

Задача фонемного транскрибирования заключается в построении последовательности фонем, отражающих произношение слова

$$\{\varphi_0\varphi_1...\varphi_n\} = f(c_0c_1...c_n), \qquad (2)$$

где $\varphi_k \in \Phi$, $c_k \in C$, Φ – множество (алфавит) фонем, C – множество (алфавит) буквенных значений, f – алгоритм транскрибирования.

Кратко рассмотрим методы фонемного транскрибирования, используемые при решении задач синтеза речи и автоматического перевода. При фонемном транскрибировании на основе правил [14], составленных вручную, необходима подготовка словарей экспертом. Алгоритм получения фонемного представления слов с использованием конечного автомата [15] строится на основе системы правил, каждое из которых преобразуется в цепочку переходов. Использование скрытых марковских моделей для фонемного транскрибирования [16, 17] предполагает этап обучения, который требует дополнительное время и большое количество данных.

В данной работе предлагается следующий метод автоматического фонемного транскрибирования.

Будем полагать, что существует статистическая зависимость между соседними буквенными значениями и, следовательно, соседними фонемами транскрибируемого слова. Тогда совместная веро-

ятность появления последовательности буквенных значений $\{c_0...c_n\}$ равна:

$$P(c_0...c_n) = P(c_0)\prod_{k=1}^{n} P(c_k \mid c_{k-1}),\ c_k \in C. \quad (3)$$

Совместная вероятность соответствующей последовательности фонем $\{\varphi_0...\varphi_n\}$:

$$P(\varphi_0...\varphi_n) = P(\varphi_0)\prod_{k=1}^{n} P(\varphi_k \mid \varphi_{k-1}),\ \varphi_k \in \Phi. \quad (4)$$

В соответствии с формулой обратной вероятности, многомерная апостериорная вероятность последовательности фонем равна:

$$P^{ac}(\varphi_0...\varphi_n) = P(\varphi_0...\varphi_n \mid c_0...c_n) = \frac{P(\varphi_0...\varphi_n)\cdot F(c_0...c_n)}{P(c_0...c_n)} \quad (5)$$

,

где $F(c_0...c_n) = P(c_0...c_n \mid \varphi_0...\varphi_n)$ – функция правдоподобия.

Предполагая независимость функций правдоподобия $F(c_k)$ и $F(c_{k-1})$, запишем

$$F(c_0...c_n) = \prod_{k=0}^{n} F(c_k). \quad (6)$$

В соответствии с критерием максимума апостериорной вероятности, результатом транскрибирования является последовательность фонем $\{\varphi_0...\varphi_n\}$ такая, что

$$P^{ac}(\varphi_0...\varphi_n) = \frac{P(\varphi_0)\prod_{k=1}^{n} P(\varphi_k \mid \varphi_{k-1})\cdot \prod_{k=0}^{n} F(c_k)}{P(c_0)\prod_{k=1}^{n} P(c_k \mid c_{k-1})} \to \max \quad (7)$$

.

В процессе поиска осуществляется сравнение фонемных представлений документа и слов запроса пользователя. В качестве меры близости фонем использовано нормированное расстояние в евклидовой метрике (8) [18].

Рассмотрим распределения $P(\varphi_i \mid c_k)$ для всех буквенных значений $c_k \in C$ и фонем $\varphi_k \in \Phi$. Нормированное расстояние в евклидо-

вой метрике между фонемами φ_i и φ_j может быть выражено формулой

$$\lambda_{\varphi_i,\varphi_j} = \sqrt{\sum_{k=1}^{n} \frac{\left(P(\varphi_i \mid c_k) - P(\varphi_j \mid c_k)\right)^2}{\sigma_k^2}}, \quad (8)$$

где σ_k^2 – несмещенная оценка среднеквадратического отклонения. Дополнительно выполняется нормирование, чтобы значение расстояния между фонемами принадлежало отрезку $[0; 1]$

$$\lambda'_{\varphi_i,\varphi_j} = 1 - \frac{\lambda_{\varphi_i,\varphi_j}}{t}, \quad (9)$$

где $t = \max\left(\lambda_{\varphi_i,\varphi_j}\right), \forall\ \varphi_i, \varphi_j \in \Phi$.

Оценка релевантности

Оценка релевантности определяется как значение меры TF-IDF, формула вычисления которой модифицирована для поиска речевых документов и основана на нахождении длины наибольшей общей подстроки [9].

Мера TF-IDF показывает важность слова для характеристики содержания конкретного документа по сравнению с другими документами коллекции. Значение данной меры пропорционально частоте вхождения слова в содержание документа и обратно пропорционально частоте вхождения этого слова в содержание других документов коллекции. При распознавании речевых документов происходит большое количество ошибок, которые приводят к искажению текста, соответствующего содержанию документа. Наличие ошибок распознавания пауз между словами не позволяет вычислять частоту слов как отношение количества вхождений слова в документ к общему числу слов документа.

Частоту TF слова запроса для речевого документа можно определить через наибольшее количество последовательно совпадающих букв (фонем) в буквенном (фонемном) представлениях фразы W_{d_k} и слова q_j

$$tf_{d_k,q_j} = \frac{L\left(W_{d_k}, q_j\right)}{l_{q_j}}, \quad (10)$$

где L – функция, которая вычисляет длину наибольшей общей подстроки между W_{d_k} и q_j, представленных в буквенном или фонемном виде, l_{q_j} – длина запросного слова q_j.

Важность слова q_j для характеристики содержания речевого документа по сравнению с другими документами коллекции можно определить через обратную частоту слова (IDF). Значение IDF [1] определяется выражением

$$idf_{q_j} = \log\left(\frac{N}{\sum_{k=1}^{N} tf_{d_k,q_j}}\right), \qquad (11)$$

где N – количество документов в коллекции.

Тогда значение оценки релевантности равно

$$F(d_k, Q) = \frac{\sum_{j=1}^{M} tf_{d_k,q_j} \cdot idf_{q_j}}{M}, \qquad (12)$$

где M – количество слов в запросе.

Эксперимент

На основе описанной модели информационного поиска речевых документов разработана SDR-система, в которой для распознавания речи используется библиотека pocketsphinx 0.8 [19], акустическая и языковая модели [20].

Таблица 1.

Значения *R*-точности

Оценка релевантности на основе TF		**Оценка релевантности на основе TF-IDF**	
буквенное представление	**фонемное представление**	**буквенное представление**	**фонемное представление**
70,86	76,93	73,66	80,26

Для проверки эффективности описанной модели проведен эксперимент, который заключался в выполнении поиска по коллекции [21], состоящей из 620 речевых документов. Эффективность поиска оценивалась значением показателя *R*-точности [22], вычисленного и усредненного по 250 запросам. Указанные запросы составлены на основе содержания речевых документов. Результаты эксперимента представлены в табл. 1.

Анализ полученных результатов (табл.1) показывает, что:

1) фонемное представление распознанного содержания речевых документов и текстового запроса позволяет повысить значение показателя R-точности при поиске по коллекции [21] в среднем на 8,5%.

2) учет значения IDF при вычислении оценки релевантности речевого документа текстовому запросу позволяет повысить значение показателя R-точности при поиске по коллекции [21] в среднем на 4%.

Вывод

Описанная модель информационного поиска речевых документов по текстовому запросу позволяет повысить показатели эффективности SDR-систем и может быть рекомендована для применения в информационных системах, содержащих мультимедийные документы.

Литература

1. *Wechsler M.* New Approaches to Spoken Document Retrieval, Information Retrieval // Information Retrieval, 2000 – Volume 3. – pp. 173-188.

2. *Jones G., Foote J., Jones K.S., Young S.* Video mail retrieval using voice: An overview of the stage-2 system. // In: van Rijsbergen C, Ed., Proceedings of the Final Workshop on Multimedia Information Retrieval (MIRO'95), Electronic Workshops in Computing – Springer. Glasgow. – 1995.

3. *Wactlar H., Hauptmann A., Witbrock M.* Informedia: News-on-demand experiments in speech recognition // In: Proceedings of DARPA Speech Recogition Workshop. – Arden House, Harriman, NY. – 1996.

4. *Wechsler M.* Spoken Document Retrieval Based on Phoneme Recognition. PhD Thesis. – ETH Zurich. – Diss. No. 12879. – 1998.

5. *Ng K., Zue V.W.* Phonetic Recognition for spoken document retrieval // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. – Vol. 1. – 1998. – P. 325-328.

6. *Witbrock M., Hauptmann A.G.* Speech recognition and information retrieval: Experiments in retrieving spoken documents // In: Proceedings of the DARPA Speech Recognition Workshop – Chantilly Virginia. – 1997.

7. *Brown M., Foote J., Jones G., Jones K.S., Young S.* Open-vocabulary speech indexing for voice and video mail retrieval // In: ACM Multimedia Conference. – Boston, MA. – 1996.

8. *Маннинг К.Д., Рагхаван П., Шютце Х.* Введение в информационный поиск // М.-Спб.-К.: изд-во Вильямс, 2011. – 520 с.

9. *Прозоров Д.Е., Яшина А.Г.* Анализ алгоритмов фонемного транскрибирования в задачах контекстного поиска речевых документов // Инфокоммуникационные технологии – Самара, 2013 – Том 12. – № 4 – С. 62-65.

10. *Glavitsch U.* The First Approach to Speech Retrieval // 1995. – http://e-collection.library.ethz.ch/eserv/eth:3328/eth-3328-01.pdf

11. Яшина А.Г. Алгоритм контекстного поиска речевых аудио-файлов на основе фонемного сравнения слов // Advanced Science, 2012. – №1. – С. 73-85. – URL: http://www.vyatsu.ru/uploads/file/ 1210/1_(2).pdf

12. *Ng K., Zue V.W.* Subword unit representations for spoken document retrieval // in Proc. Eur. Conf. Speech Communication Technology – 1997.

13. Snowball // [Электронный ресурс] – URL: http://snowball.tartarus.org /algorithms/russian/stemmer.html

14. *Логачева В.К., Клышинский, Галактионов В.А.* Современные методы практической транскрипции // Препринты ИПМ им. М.В.Келдыша. 2012. – № 13. – 18 с. – URL: http://library.keldysh.ru/preprint.asp?id=2012-13

15. *Логачева В.К., Клышинский Э.С., Галактионов В.А.* Автоматическая генерация правил транскрипции и машинная транскрипция имен собственных с использованием конечного автомата // Препринты ИПМ им. М.В.Келдыша. 2012. – № 14. – 24 с. – URL: http://library.keldysh.ru/preprint.asp?id=2012-14

16. *Яшина А.Г.* Поиск речевых документов на основе фонемного транскрибирования слов с использованием скрытой марковской модели // Всероссийская ежегодная научно-техническая конференция «Общество, наука, инновации» (НТК-2013): 15-26 апр. 2013 г. : сб. материалов / Вят. гос.ун-т ; отв. ред. С. Г. Литвинец. – Киров, 2013.

17. *Taylor P.* Hidden Markov Models for Grapheme to Phoneme Conversion // In Proceedings of INTERSPEECH – 2005.

18. *Яшина А.Г.* Поиск речевых документов с использованием различных мер сравнения фонем // Сб. Трудов МНТК «Актуальные направления фундаментальных и прикладных исследований» – М, 2013. – Том 2. – С. 73-76.

19. CMU Sphinx. Open Source Toolkit For Speech Recognition // http://cmusphinx.sourceforge.net

20. Voxforge-ru-0.2 // URL: http://sourceforge.net/projects/cmusphinx/files /Acoustic%20and%20Language%20Models/Russian%20Voxforge/

21. FestLang // URL: http://sourceforge.net/projects/festlang.berlios

22. *Агеев М., Кураленок И., Некрестьянов И.* Официальные метрики РОМИП 2010 // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2010. – Казань, 2010.