Исследование характеристик системы поиска ключевых слов на основе минимального интервала редактирования и мер доверительности

Янь Цзинбинь,
аспирант
И.Э. Хейдоров,
доцент, кандидат физико-математических наук
А.В. Ткаченя,
студент

В данной работе для поиска ключевых слов предлагается использовать двухэтапный алгоритм на основе решётки слогов и минимального интервала редактирования для детектирования временных координат возможных ключевых слов и верификации найденных слов на основе мер доверительности. Эксперимент показал, что использование решётки слогов обеспечивает большую точность распознавания, чем решётка на основе фонем. Точность поиска ключевых слов при этом составляет 88.2%. Использование метода опорных векторов (МОВ) для объединения мер доверительности позволило уменьшить вероятность ложной тревоги до 8.8%, что позволяет на этой основе создавать системы поиска ключевых слов с приемлемыми с практической точки зрения характеристиками.

Введение

Поиск ключевых слов в речевых файлах является одной из наиболее сложных задач в области обработки речи. С её помощью можно реализовать аудиоиндексацию и поиск информации, например: поиск информации в Интернете, контроль речевой связи, управление речевыми библиотеками [1, 2].

Наиболее простой метод поиска ключевых слов использует распознаватель с большим словарём для перевода непрерывной речи в текст. Для поиска ключевого слова осуществляется поиск в полученном тексте с использованием традиционных алгоритмов поиска текста. Проблема этого метода состоит в том, что из-за ограниченного множества слов в распознавателе невозможно распознать слова, отсутствующие в словаре, например: имена, акронимы и слова из иностранных языков.

Другой метод поиска ключевых слов основан на скрытых Марковских моделях (СММ). Он использует СММ для каждого ключевого слова и одну «модель мусора» для всех остальных слов [3]. Этот метод не имеет ограничений при условии, что определено множество ключевых слов, которые необходимо найти. Но для каждого нового ключевого слова необходимо не только обучать новую СММ-модель, но также нужно заново обучать «модель мусора». Поэтому использование этого метода при определённых условиях является сложным.

На сегодняшний день наиболее популярным решением задачи поиска ключевых слов в потоке слитной речи является использование акустико-фонетической СММ и вычисление апостериорных вероятностей фонемной решётки, каждый узел которой ассоциируется с моментом времени в рамках произнесённой речи. Данный метод обладает большой гибкостью, и результат поиска не зависит от словарного множества распознавателя, что позволяет осуществлять поиск для любого запрашиваемого ключевого слова без дополнительного обучения системы. Однако данный подход сопряжён со значительной вычислительной сложностью и, следовательно, требует введения некоторых дополнительных процедур для создания систем поиска ключевых слов в реальном масштабе времени. Кроме того, отсутствие в явном виде словаря требует введения дополнительной процедуры верификации найденных ключевых слов.

В связи с этим в данной работе для поиска ключевых слов предлагается использовать двухэтапный алгоритм на основе решётки слогов и минимального интервала редактирования (МИР) для детектирования временных координат возможных ключевых слов и верификации найденных слов на основе мер доверительности.

Структура системы поиска ключевых слов на основе решётки слогов

Для осуществления поиска и верификации ключевых слов была предложена следующая схема системы ***(рис. 1)***. После выделения признаков речевого сигнала последовательность наблюдений $O=\{o_1,o_2,...,o_T\}$, где T — количество векторов-признаков, поступает на вход системы распознавания,

созданной на основе СММ, которая позволяет преобразовать последовательность наблюдений в последовательность некоторых структурных единиц речи (фонем, аллофонов и т.д.). Для вероятностей акустических единиц речи, полученных на выходе СММ, строим решётку $L = (N, A, n_{start}, n_{end})$ — направленный непериодический граф, где N — множество узлов, A — множество связей между узлами и n_{start}, $n_{end} \in N$ — начальный и конечный узлы решётки соответственно ***(рис. 2)***. Связь представляется в виде $a = (S[a], E[a], I[a], w[a])$, где $S[a]$, $E[a] \in N$ — начальный и конечный узлы; $I[a]$ — фрагмент речи (слог или фонема); $w[a] = p_{ac}(a)^{1/\lambda}$ — весовой коэффициент связи, который представляет собой вероятность перехода между узлами; $p_{ac}(a)$ — акустическое сходство; λ — весовой коэффициент.

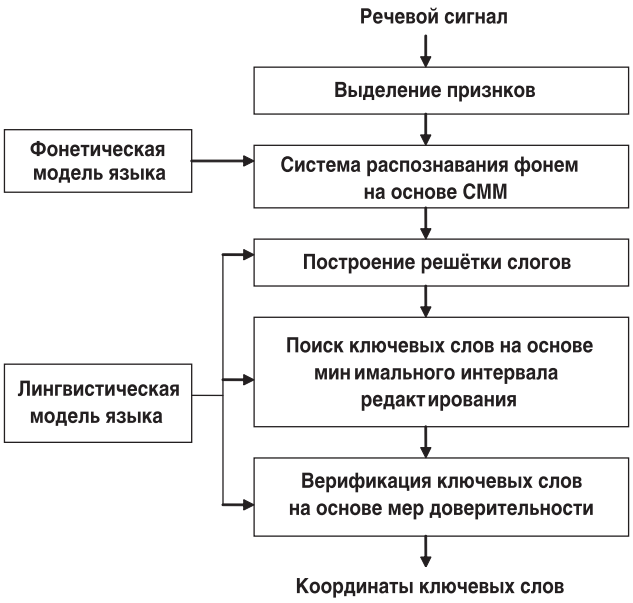

Рис. 1. Структура системы поиска ключевых слов

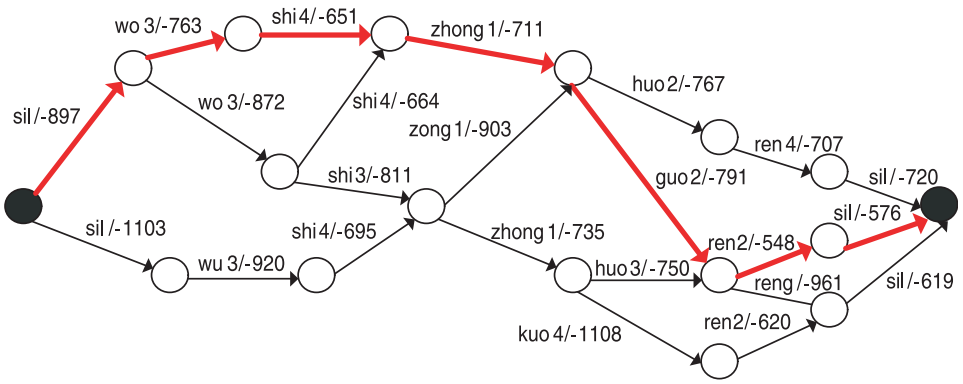

Рис. 2. Решетка слогов речи

Алгоритм минимального интервала редактирования (МИР) [4]

Для последовательности слогов, полученной на выходе решётки, характерно наличие трёх основных типов ошибок. Ошибка типа «замена» отвечает случаю, когда вместо одного слога система распознала другой. «Вставка» — это появление в потоке слогов, которых не было в исходной речи. И, наконец, ошибка типа «удаление» характерна случаю пропуска слога, присутствующего в реальной речи. Для повышения эффективности поиска ключевых слов в систему необходимо ввести дополнительный модуль, позволяющий учесть априорную информацию об ошибках такого рода.

Определим понятие минимального интервала редактирования между строками U и V как минимальные затраты на преобразование строки U в строку V с помощью трёх основных операций: вставки, удаления и замены. Для хранения значений минимального расстояния используем матрицу $M(0,...,p)(0,...,q)$, где p и q — это длины строк U и V соответственно. Элементы матрицы M для строк U и V вычисляются следующим образом:

$$M(0)(0) = 0;$$
$$M(i)(0) = i * I; i = 1,..., p;$$
$$M(0)(j) = j * D; j = 1,..., q;$$
$$M(i)(j) = \min\{M(i-1)(j-1) + S(U(i), V(j)), M(i-1)(j) + D, M(i)(j-1) + I\};$$

где S, I и D — затраты на операции замены, вставки и удаления соответственно.

Пусть $K=\{k_1, ..., k_N\}$ — это последовательность слогов ключевого слова, которое нужно найти в слоговой решётке; $Q=\{q_1, ..., q_M\}$ — последовательность слогов предполагаемого ключевого слова; C_S, C_I, C_D — функции затрат для замены, вставки и удаления; $MED(K,Q,C_S,C_I,C_D)$ — функция минимального расстояния, которая в качестве своего значения возвращает элемент $M(N,M)$; S_{max} — пороговое значение для функции MED.

Алгоритм минимального расстояния реализуется следующим образом.
1) Получение предполагаемых ключевых слов.
2) Для каждого предполагаемого ключевого слова:
a) получение последовательности слогов $Q=\{q_1, ..., q_M\}$;
b) вычисление минимальной корректировки $S=MED(K,Q,C_S,C_I,C_D)$;
c) если $S \le S_{max}$, тогда Q есть ключевое слово.

Использование данного алгоритма позволяет определить временные координаты возможных ключевых слов с учётом всех возможных ошибок, совершённых на уровне фонетического распознавания. Однако за счёт этого данный подход имеет большое количество ложных тревог, когда алгоритм принимает решение о наличии ключевого слова, хотя в данный момент времени оно отсутствует. В связи с этим в работе предлагается ввести дополнительный этап верификации ключевых слов, основанный на использовании мер доверительности [5].

Метод доверительного интервала

Представим речевой сигнал на входе системы поиска ключевых слов в виде последовательности наблюдений $O=\{o_1,o_2,...,o_T\}$. Обозначим модель ключевого слова как Λ и определим меру доверительности как количественную оценку совпадения O и Λ. Другими словами, мера доверительности определяется вероятностью генерации последовательности O на основе модели Λ. Для моделирования речевых сигналов на основе СММ компоненты вектора признаков (вектора наблюдений) предполагаются независимыми, а для каждого состояния плотность распределения вероятностей наблюдений представляется гауссовыми смесями (ГС), описывающими вероятность вектора наблюдений o_t в состоянии j как

$$b_j(o_t) = \prod_{s=1}^{S}\left[\sum_{m=1}^{M_s} c_{jsm} N\left(o_{st}; \mu_{jsm}; \sigma_{jsm}\right)\right],$$

где S — размерность вектора признаков речевого сигнала; M_s — размерность ГС для компоненты s; c_{jsm}, μ_{jsm}, σ_{jsm} — соответственно весовое значение, среднее значение и дисперсия для m-ой компоненты ГС, по предположению представляющей собой нормальное распределение:

$$N\left(o_{st}; \mu_{jsm}; \sigma_{jsm}\right) = \frac{1}{\sqrt{2\pi}\sigma_{jsm}} \exp\left(-\left(o_{st} - \mu_{jsm}\right)^2 / 2\sigma_{jsm}^{\ 2}\right).$$

Пусть j-ое состояние модели Λ представляется подпоследовательностью $O^k = \{o_k, o_{k+1}, ..., o_K\}$. Тогда определим значение меры доверительности CM_j следующим образом:

$$CM_j = \frac{\sum_{t=k}^{K}\sum_{m=1}^{M_s}\sum_{s=1}^{S} D\left(o_{st}; \mu_{jsm}, \sigma_{jsm}\right)}{(K-k)M_s S},$$

где

$$D(o_{ti}; \mu_{jsm}, \sigma_{jsm}) = \begin{cases} 1 & (o_{st} - \mu_{jsm}) \in [\mu_{jsm} - k\sigma_{jsm}, \mu_{jsm} + k\sigma_{jsm}] \\ 0 & \textit{иначе} \end{cases},$$

где k — управляющий параметр для доверительного интервала. Определим меру доверительности для каждого ключевого слова Λ путём нормализации значений для каждого состояния:

$$CM_1 = \frac{1}{N}\sum_{j=1}^{N} CM_j,$$

где N — количество состояний СММ. Верификация ключевого слова происходит путём сравнения полученной меры доверительности с некоторым порогом, выбранным, как правило, эмпирически.

Данная мера доверительности обеспечивает достаточно высокую точность верификации и имеет большой потенциал по улучшению характеристик. Алгоритм расчёта меры доверительности на основе использования доверительного интервала подразумевает получение множества промежуточных результатов, использование которых позволит повысить точность при незначительном увеличении вычислительной сложности.

Акустическая мера доверительности на основе нормализации длительности состояния

Одним из недостатков представленной выше меры доверительности является отсутствие её нормализации на длину состояния СММ, вследствие чего возможны ситуации, когда состояние с малой длительностью затеняет результаты для более длительной последовательности наблюдений.

Обозначим СММ как $\lambda = \{N, \pi, A, B\}$, где N — число состояний СММ $S = \{S_1, S_2, \ldots S_N\}$, π — матрица начальных вероятностей, A и B — матрицы переходных вероятностей и вероятностей наблюдения соответственно. Обозначим начальный и конечный моменты времени нахождения системы в состоянии i как $b[i]$ и $e[i]$. Тогда определим нормализованную меру доверительности следующим образом:

$$CM_2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]}\log p(o_t \mid s_i)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]}\log b_i(o_t)\right]$$

Путём замены $\log p(o_t \mid s_i)$ на $\log(s_i \mid o_t)$ согласно формуле Байеса получим ещё одно выражение для меры доверительности:

$$CM_3 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]}\log p(s_i \mid o_t)\right],$$

где

$$p(s_i \mid o_t) = \frac{p(o_t \mid s_i)p(s_i)}{\sum_{j=1}^{N} p(o_t \mid s_j)p(s_j)} = \frac{b_i(o_t)\,p(s_i)}{\sum_{j=1}^{N} b_j(o_t)\,p(s_j)} \quad .$$

Обе эти меры доверительности представляют собой среднее значение акустической вероятности в рамках СММ.

Мера доверительности на основе динамического рейтинга

Описанные выше методы подтверждения ключевых слов в той или иной степени используют модели фонем. Они просты и эффективны, в значительной степени позволяют снизить вероятность ложной тревоги. Рассмотрим ещё один метод верификации ключевых слов, основанный на использовании апостериорной сети и динамического рейтинга.

В результате работы декодирующего алгоритма Витерби текущему вектору наблюдений o_t входящего речевого сигнала и каждой допустимой в данный момент модели слова Λ_j ставится в соответствие последовательность состояний модели $S(\Lambda_j, o_t)$. Определим меру соответствия (характеристическое значение) $L_j(o_t)$ последовательности состояний $S(\Lambda_j, o_t)$ модели Λ_j в момент времени t следующим образом:

$$L_j(o_t) = \ln P\left(o_t \mid S\left(\Lambda_j, o_t\right)\right) \quad j = 1,2,\ldots, N(o_t) \qquad (1)$$

где $N(o_t)$ — число активных моделей в момент времени t. Отсортируем набор характеристических значений по убыванию для всех моделей:

$$L_{j_1}(o_t) > L_{j_2}(o_t) > \cdots > L_{j_k}(o_t) > \cdots > L_{j_{N(o_t)}}(o_t) \ .$$

Пусть характеристическое значение $L_{Kw}(o_t) = L_{jk}(o_t)$ для модели ключевого слова Λ_{Kw} занимает в рейтинге k-ое место, тогда определим динамический рейтинг на o_t-ом фрейме как $k/N(o_t)$:

$$Q(o_t \mid \Lambda_{Kw}) = \frac{\sum_{k=1}^{N(o_t)} G(L_k(o_t) - L_{Kw}(o_t))}{N(o_t)} \ ,$$

где

$$G(L_k(o_t) - L_{Kw}(o_t)) = \begin{cases} 0 & \text{if } L_k(o_t) \le L_{Kw}(o_t) \\ 1 & \textit{иначе} \end{cases} \ .$$

На основе вышеприведённых рассуждений представим динамический рейтинг как обобщённое характеристическое значение в рамках всей анализируемой длительности сигнала:

$$CM_4(O \mid \Lambda_{Kw}) = \frac{1}{T}\sum_{t=1}^{T} Q(o_t \mid \Lambda_{Kw}) \ .$$

Для верификации ключевого слова необходимо полученное значение динамического рейтинга сравнить с порогом. Особенностью метода динамического рейтинга является то, что порог отторжения может быть установлен одинаковым для всех ключевых слов.

Следует отметить, что вычисление динамического рейтинга для ключевых слов не приводит к существенному увеличению вычислительной сложности по сравнению с алгоритмами на основе акустической меры доверительности. В процессе декодирования Витерби значения вероятностей для выражения (1) уже рассчитаны, поэтому вычисление $Q(o_t)$ и нормализация приводят к дополнительным KT сложениям и вычитаниям, а также T+1 делениям.

Однако данный алгоритм обладает рядом существенных преимуществ. Во-первых, верификация ключевых слов на основе динамического рейтинга имеет более стабильный характер, особенно в условиях шума. Имеющийся в речевых данных шум влияет на абсолютные значения акустических вероятностей, однако не оказывает практически никакого влияния на место в рейтинге, а соответственно, и на характеристическое значение.

Во-вторых, динамический рейтинг рассчитывается исключительно на основании значений акустической вероятности, поэтому изменение словаря ключевых слов не оказывает никакого влияния на работоспособность алгоритма.

В-третьих, как уже упоминалось ранее, порог отторжения может быть установлен единым для всех ключевых слов, а кроме того, наличие шума в исходных данных не влияет на абсолютное значение порога.

Представленные выше меры доверительности позволяют провести верификацию найденных ключевых слов. Особенно хороших результатов позволяет добиться использование комплексных критериев верификации, построенных на основе использования нескольких мер доверительности одновременно.

В данной статье для объединения мер доверительности и проведения экспериментов был использован МОВ [6].

Эксперимент

Для проведения эксперимента и определения сравнительных характеристик предложенной системы поиска ключевых слов была использована база данных, содержащая 124 часа речи. С использованием этой базы данных для создания акустико-фонетической модели речи были обучены СММ, в качестве вектора признаков был использован вектор из 39-ти мел-кепстральных характеристик и их производных. Тестирование системы производилось с использованием реальной слитной речи продолжительностью 3,54 часа, в которой обнаружению подлежали 13 часто повторяемых ключевых слов, появившихся в тестовых фрагментах 794 раза.

Для оценки характеристик системы поиска ключевых слов были использованы в качестве основных следующие величины: вероятность обнаружения Pd (процент правильно найденных ключевых слов), вероятность ложного отказа (процент неправильно отторгнутых ключевых слов), вероятность ложной тревоги FAR (процент принятия ложных слов в качестве ключевых). Для представления фонем была выбрана непрерывная СММ с пятью состояниями, каждое из которых моделировалось ГС.

Первый эксперимент был посвящён выбору базовой единицы распознавания и влиянию процедуры построения гипотез на основе МИР на результат правильного обнаружения ключевых слов ***(см. табл. 1)***.

Таблица 1

Вероятность правильного обнаружения ключевых слов в зависимости от структурной единицы решетки

Структурная единица решётки	Вероятность обнаружения Pd без использования МИР, %	Вероятность обнаружения Pd с использованием МИР, %
Фонема	80.7%	82.7%
Слог	85.2%	88.2%

Как видно из таблицы 1, использование слогов в качестве минимальных структурных единиц речи для задачи поиска ключевых слов является более предпочтительным. Кроме того, использование алгоритма МИР для формирования последовательности слогов с учётом возможных ошибок типа «замещение», «вставка» и «пропуск» позволило увеличить вероятность правильного обнаружения ключевых слов в среднем на 3%.

Второй эксперимент был посвящён тестированию точности верификации ключевых слов в зависимости от используемой меры доверительности ***(см. табл. 2)***.

Из таблицы 2 видно, что МОВ, объединяющий несколько различных мер доверительности, позволяет достичь намного лучших характеристик отторжения, чем каждая из представленных мер доверительности в отдельности. Изменение параметров МОВ позволяет в определённых пределах изменять вероятность ложной тревоги, настраивая систему под конкретные условия.

Таблица 2

Вероятность ложной тревоги системы верификации ключевых слов в зависимости от меры доверительности

FAR без верификации, %	FAR для меры динамического рейтинга, %	FAR для доверительного интервала, %	FAR для комплексного алгоритма МОВ, %
40.5	22.5	30.2	8.8

Эксперимент показал, что предложенный метод на основе МОВ, объединяющей различные меры доверительности, позволяет достичь точности верификации ключевых слов более высокой, чем при использовании этих мер доверительности по отдельности.

Использование МОВ для объединения мер доверительности позволило уменьшить вероятность ложной тревоги до 8.8% при сохранении той же точности правильного обнаружения, что позволяет на этой основе создавать системы поиска ключевых слов с приемлемыми с практической точки зрения характеристиками.

Заключение

В данной работе предложена двухэтапная структура системы поиска ключевых слов на основе решётки слогов с использованием алгоритма минимального расстояния и верификации найденных слов с использованием мер доверительности.

Результат эксперимента показал, что решётка слогов позволяет достичь большей точности, чем решётка фонем, и при этом точность поиска ключевых слов составляет 88,2%.

Использование МОВ для объединения мер доверительности позволило уменьшить вероятность ложной тревоги до 8,8% при сохранении той же точности правильного обнаружения, что позволяет на этой основе создавать системы поиска ключевых слов с приемлемыми с практической точки зрения характеристиками.

Литература

1. *J.Mamou, D.Carmel, R.Hoory.* Spoken Document Retrieval from Call-Center Conversations. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA. — 2006. — P. 51–58.

2. *Chung-Hsien Wu, Yeou-Jiunn Chen.* Multi-keyword Spotting of Telephone Speech Using a Fuzzy Search and Keyword-driven Two-level CBSM. Speech Communication. — 2001, — 33(3). — P. 197–212.

3. *Rose R.C.* Keyword detection in conversational speech utterance using hidden Markov model based continuous speech recognition. Computer Speech and Language. — 1995. — vol. 9. — P. 309–333.

4. *A. Kartik, V. Ashish.* Keyword Search Using Modified Minimum Edit Distance Measure. IEEE International Conference on Acoustics, Speech and Signal Processing. — 2007. — vol. 4. — P. 929–932.

5. *Mingxing Xu, Fang Zheng, Wenhu Wu, Ditang Fang.* Research on rejection method for continuous speech keyword spotting system. Journal of Tsinghua University. — 1998. — vol. 38(1). — P. 89–91.

6. *Vapnik, V.N.* Statistical Learning Theory. New York, Wiley, 1998.

Янь Цзинбинь —

аспирант БГУ, научные интересы — методы и алгоритмы обработки речевых сигналов, теория метода опорных векторов, алгоритмы обнаружения ключевых слов в потоке речи.

Хейдоров Игорь Эдуардович —

Окончил с отличием БГУ (Минск) в 1996 году, с 1998 года работает на кафедре радиофизики БГУ, к.ф.-м.н. (2000 г), доцент. Сфера научных интересов — методы и алгоритмы распознавания и синтеза речи, автоматическая индексация аудиодокументов. Автор 40 работ.

Ткаченя А.В. —

студент БГУ, факультет радиофизики и электроники . Область научных интересов — системы анализа и индексирования аудиосигналов, скрытые Марковские модели в задачах распознавания речи.