УДК 934.1; 272

ПРОБЛЕМЫ РЕАЛИЗАЦИИ РАСПОЗНАВАТЕЛЯ РЕЧИ НА РЕКУРРЕНТНОМ ПРОЦЕССОРЕ

Рождественский Ю.В., Дьяченко Ю.Г., Морозов Н.В.

Введение

Распознаватель речи является необходимым компонентом интеллектуальных диалоговых систем, начинающих широко внедряться в бытовую технику и постепенно становящихся необходимой составляющей среды обитания человека ("умный дом", "умный автомобиль" и т.д.). Реализация распознавателя речи предъявляет довольно высокие требования к производительности вычислителя, на котором он построен: цифрового сигнального процессора (ЦСП) с традиционной или параллельной архитектурой [1]. Однако оно наталкивается на проблему необходимости сокращения требуемых вычислительных ресурсов до уровня, обеспечиваемого типовыми ЦСП. К наиболее важным вычислительным ресурсам относятся производительность ЦСП и объем памяти данных в нем. Как правило, величина этих характеристик определяет и стоимость процессора, используемого для реализации распознавателя речи.

Данная работа посвящена решению актуальных проблем реализации распознавателя речи на ЦСП низкой и средней производительности. Предлагаются алгоритмы, обеспечивающие упрощение реализации распознавателя речи с точки зрения требуемых вычислительных ресурсов при сохранении высокого качественного уровня распознавания, отвечающего современным требованиям (для словаря из 100 слов точность распознавания 95% в фоновом "белом" шуме с SNR=15 дБ, соответствующем уровню уличного шума).

Полученные результаты и учет фундаментальной параллельности [2] при обработке звуковой информации служат основой для разработки архитектуры параллельного динамического (рекуррентного) речевого процессора (РРП) [3], максимально адаптированной к решению задач обработки речевой информации.

Системы распознавания речи

По своей сути процесс распознавания речи представляет собой процесс сравнения входного словарного потока, произносимого на выбранном языке, с моделью данного языка (словарь + правила образования предложений). Системы распознавания речи классифицируются по типу входного речевого потока (распознаватели изолированных слов и распознаватели слитной речи) и по степени открытости (зависимые от диктора и независимые от диктора распознаватели).

Наиболее сложными для реализации являются независимые от диктора распознаватели слитной речи. Они требуют больших вычислительных ресурсов и огромного объема памяти для хранения базы данных, включающей в себя как модели речевых фрагментов (слогов, слов и т.д.), так и разнообразные правила (грамматические, синтаксические, семантические и т.д.), позволяющие правильно сегментировать входной речевой поток и идентифицировать отдельные слова и фразы. Как правило, такие системы

распознавания речи реализуются на мощных, высокопроизводительных вычислительных комплексах на базе ПЭВМ.

Наименее сложными для реализации являются настроенные на ограниченный круг дикторов распознаватели изолированных слов. Они требуют меньших вычислительных ресурсов, так как модели слов оказываются менее объемные, а применение правил, позволяющих сегментировать входной речевой поток на слова и фразы, не требуется. Процесс распознавания изолированных слов-команд заключается в сравнении и выборе наиболее близкого по произношению слова. При этом каждое слово хранится в словаре в виде некоторой реализации звуковой модели.

Одними из наиболее широко используемых в настоящее время являются системы распознавания последовательности изолированных слов, независимые от диктора. Они реализуют методологию распознавания изолированных слов и по требуемым вычислительным ресурсам занимают промежуточное положение между предыдущими двумя системами. Реализация именно таких систем на РРП рассматривается в данной работе.

Наиболее часто для описания модели слова используются дискретные скрытые марковские модели (СММ) [1]. Один из вариантов СММ, обеспечивающий наилучшее качество распознавания слова, приведен на рис.1. Это модель первого порядка, близкая к традиционной СММ.

Модель слова состоит из N узлов (состояний), S_i, отображающих растянутое во времени произнесение моделируемого слова. Каждое состояние модели, за исключением последнего, характеризуется двумя возможными переходами, отражающими движение процесса произнесения слова по описывающей его модели (рекурсия и переход к следующему по порядку узлу), которые характеризуются своими вероятностями, $a_{i,j}$.. Возвраты к предыдущим узлам и переходы к узлам S_{i+k}, минуя узел S_{i+1},. запрещены.

Переходы сопровождаются наблюдениями O_t, которые регистрируются в это время. Каждое наблюдение – это характеристический вектор, выделяемый на соответствующем временном интервале (фрейме) из произнесения слова, прошедший процедуру квантования на векторной кодовой книге (ВКК). Каждому j-му кодовому вектору из кодовой книги векторов соответствует вероятность его наблюдения во время всех переходов модели слова, $b_{i,k}(j)$. Значение $b_{i,k}(j)$ представляет собой вероятность появления кодового вектора с номером j из кодовой книги векторов при переходе процесса из состояния S_i в состояние S_k.

Одно и то же слово имеет различное звучание, зависящее как от акустических особенностей голосового тракта, так и от эмоционального состояния и особенностей характера говорящего. Точный математический учет всего спектра факторов, влияющих на особенности произнесения слова, пока не представляется возможным. Поэтому модели слов носят статистический характер, а задача их создания требует обработки большого статистического материала – тренировки, или обучения.

Традиционно система распознавания изолированных слов включает в себя две подсистемы (рис.2):

- *речевого обучения*, формирующую ВКК и библиотеку СММ для распознавателя на основе речевой базы (РБ). ВКК тренируется на всей РБ, в то время как СММ тренируются на произнесениях только слов из состава словаря, на который настроена система распознавания,
- собственно *распознавателя речи*.

При реализации распознавателя речи на ЦСП ВКК и СММ, полученные на этапе тренировки распознавателя, записываются во внутреннюю или внешнюю память и используются во время реальной работы распознавателя.

Подсистема обучения реализуется обычно на ПЭВМ, поскольку обработка

статистического материала – РБ –, требует больших вычислительных ресурсов и памяти. В исключительных случаях, при создании индивидуального распознавателя, рассчитанного на работу с одним-двумя пользователями и маленьким словарем (10-20 слов), подсистема тренировки может располагаться и на ЦСП с достаточным объемом внешней памяти. В этом случае внешняя память будет использоваться для накопления РБ и, возможно, для хранения ВКК и СММ, если они не уместятся во внутренней памяти ЦСП.

Результат распознавания – номер СММ, наилучшим образом соответствующей произнесенному слову,– используется для накопления информации обо всей фразе целиком и принятия решения о выполнении требуемой команды.

Алгоритмическая основа распознавателя

Традиционная структура алгоритма распознавания речи, обеспечивающего выполнение современных требований к качеству распознавания речи, представлена на рис.3.

Входная предобработка речевого сигнала включает в себя пре-фильтрацию оцифрованного сигнала с целью уменьшения низкочастотных и высокочастотных шумов и разбиение непрерывного потока отсчетов речевого сигнала на фреймы длительностью 10–20 мсек. Поскольку частоты реального речевого сигнала лежат в диапазоне 150 – 3500 Гц, в качестве входного фильтра можно использовать фильтр Баттерворта четвертого порядка с частотой отсечки 120 Гц. Его импульсная характеристика описывается формулой:

$$H(z) = \prod_{k=1}^{2} \frac{a_{0k} + a_{1k} z^{-1} + a_{2k} z^{-2}}{1 + b_{1k} z^{-1} + b_{2k} z^{-2}},$$

где a_{0k}, a_{1k}, a_{2k}, b_{1k} и b_{2k} — коэффициенты фильтра. Он обеспечивает хорошее подавление низкочастотных шумов и устраняет постоянное смещение уровня речевого сигнала.

Шумоподавление необходимо в силу того, что эксплуатация систем распознавания в реальных условиях обычно происходит в присутствии довольно сильного фонового шума: уличного, домашнего, производственного и т.д. Оно осуществляется в частотной области на основе методики RASTA фильтрации [4] барковского спектра текущего фрейма речевого сигнала. Традиционный подход в этом случае включает в себя следующие этапы:

1. Взвешивание пре-фильтрованного речевого сигнала окном Хемминга:

$$W_H(n) = 0.54 - 0.46\, cos(2\pi n / (L_H - 1))$$

пофреймно, где L_H – размер окна в отсчетах.

2. Вычисление энергетического спектра сигнала $P(k)$ на текущем фрейме. Традиционно наиболее эффективным методом для выполнения этой операции считается быстрое преобразование Фурье (БПФ) на 256 или 512 точек.

3. Преобразование энергетического спектра сигнала $P(k)$ в барковский спектр $A(n)$, имеющий существенно меньшую размерность [5].

4. Нелинейное преобразование барковского спектра – перевод его в логарифмический масштаб с предварительным взвешиванием:

$$A_{ln}(n) = ln\,[1 + J \cdot A(n)],\ n=N_B,$$

где J – взвешивающий коэффициент, значение которого определяется уровнями энергетики входного сигнала и шума и колеблется в диапазоне $2 \cdot 10^{-8} \div 3 \cdot 10^{-10}$; N_B – число

барковских полос.

5. Применение RASTA фильтрации к нелинейному спектру $A_{ln}(n)$. Спектр шума в реальных условиях меняется медленнее по сравнению с речью. Его можно удалить с помощью фильтра низкой частоты:

$$H_{RASTA}(z) = \frac{a_1 + a_2 z^{-1} + a_3 z^{-3} + a_4 z^{-4}}{1 + bz^{-1}}.$$

6. Возвращение фильтрованного барковского спектра к линейному масштабу:

$$A'(n) = exp\,[A'_{ln}(n)] / J.$$

Такая методика позволяет эффективно подавить как аддитивную, так и мультипликативную составляющую шума за счет относительно небольших вычислительных затрат.

В качестве *речевых параметров*, на основе которых строятся модели слов или фраз, традиционно используются коэффициенты линейного предсказания и производные от них – кепстральные параметры, обеспечивающие лучшую точность распознавания.

Квантование вектора параметров по векторной кодовой книге позволяет упростить следующие этапы – поиск ближайшей модели и принятие решения, – за счет сокращения размерности решаемой задачи. Квантование заключается в поиске вектора из кодовой книги, расстояние от которого до квантуемого вектора параметров речевого фрейма минимально. В качестве меры близости традиционно выбирается евклидово расстояние. Размерность кодовой книги составляет обычно 256 векторов.

Этап *поиска ближайшей модели* фразы состоит в определении модели из базы данных, наилучшим образом соответствующей полученной последовательности наблюдений. Среди различных методов и моделей, лежащих в основе такого поиска, наименьшей вычислительной сложностью при высоком уровне эффективности обладает метод Витерби на дискретных СММ, описанных выше. Работая в области логарифмических значений вероятностей, он позволяет обойтись без операций умножения в процессе вычислений.

Такой подход обеспечивает точность распознавания изолированных слов 95% для словаря из 100 слов в условиях фонового "белого" шума с отношением сигнал/шум (SNR) 15 дБ, соответствующем уровню уличного шума. При этом оценка требуемых вычислительных ресурсов для реализации описанного алгоритма равна примерно 20 MIPS (миллионов инструкций в секунду) и примерно 6 кбайт памяти данных для типового 16-разрядного ЦСП. Львиная доля затрат (80 – 85%) приходится на вычисление барковского спектра и на процедуру квантования характеристических векторов. Кроме того, для вычисления барковского спектра с помощью БПФ требуется значительный объем памяти данных. Возникает естественное желание модифицировать и упростить алгоритмы распознавателя с целью сокращения вычислительных затрат и требуемого объема памяти данных.

Модификация алгоритмической базы

Исследования показали, что преобразование Фурье для вычисления спектра исходного сигнала с последующим вычислением барковского спектра может быть успешно заменено банком полосовых фильтров. При этом качественные характеристики распознавателя речи практически не изменяются. Количество фильтров в банке соответствует числу барковских частотных полос [4]. Оно связано с особенностью слухового аппарата человека [5] и свойством фундаментальной параллельности,

присущей процессам восприятия и обработки звуковой информации [2], и может быть определено по формуле:

$$N_B = \left[6 \cdot \ln\left(\frac{F_S}{1200} + \sqrt{\left(\frac{F_S}{1200}\right)^2 + 1} \right) \right],$$

где F_S — частота дискретизации речевого сигнала; скобки “[]” означают усечение числа до ближайшего целого.

Уравнение каждого полосового фильтра описывается формулой:

$$\boldsymbol{H_B(z)} = \frac{\boldsymbol{c_k}}{1 + \boldsymbol{d_{1k}} z^{-1} + \boldsymbol{d_{2k}} z^{-2}} \cdot \frac{\boldsymbol{c_k}}{1 + \boldsymbol{d_{3k}} z^{-1} + \boldsymbol{d_{4k}} z^{-2}}, \quad k = 1, ..., N_B,$$

где c_k,, d_{1k},, d_{2k},, d_{3k} и d_{4k} — коэффициенты фильтра, зависящие от номера полосы барковских частот и от частоты дискретизации F_S.

В процессе фильтрации для каждой полосы подсчитывается суммарная энергия фильтрованного сигнала, значения которой и формируют спектр, аналогичный тому, который получается в результате применения преобразования Фурье и перевода энергетического спектра речевого сигнала в область барковских частот [4].

Полученные оценки энергии речевого сигнала по отдельным частотным полосам используются также для принятия решения о характере текущего речевого фрейма. Если уровень энергии хотя бы в одной из полос превышает заданный порог, текущий фрейм считается голосовым. В противном случае он рассматривается как шумовой (пауза). Вектора параметров вычисляются только для голосовых фреймов. Этот же критерий служит индикатором начала произнесения слова.

Оценки показывают, что такое упрощение позволяет достаточно точно реконструировать барковский спектр речевого сигнала. Среднее отклонение рассчитанной таким образом энергии сигнала по барковским полосам от энергии, вычисленной с помощью преобразования Фурье, не превышает 2%. Поскольку такая методика используется как на предварительном этапе тренировки моделей, так и на этапе реального распознавания, точность распознавания ухудшается примерно на 0.1%. В то же время замена преобразования Фурье банком полосовых фильтров позволяет значительно сократить требующуюся производительность ЦСП (до 12 MIPS) и объем памяти данных в ЦСП (до 3 кбайт). Кроме того, использование банка полосовых фильтров позволяет эффективно распараллелить вычисления на архитектуре РРП.

Для решения ряда практических задач эффективно используются 8-разрядные микроконтроллеры. Они обладают существенно меньшей производительностью по сравнению с 16-разрядными ЦСП, не позволяющей использовать их для непосредственной реализации распознавателя речи. Однако дальнейшее упрощение алгоритмов распознавателя позволяет решить и эту задачу за счет небольшой потери качества распознавания. Это открывает дополнительные возможности для реализации архитектуры РРП в базисе 8-разрядных процессорных элементов или 16-разрядных процессорных элементов, поддерживающих одновременное исполнение двух 8-разрядных операций.

Дальнейшее упрощение алгоритмов распознавателя речи достигается путем использования полосовых фильтров второго порядка:

$$\boldsymbol{H_B(z)} = \frac{\boldsymbol{c_k}}{1 + \boldsymbol{d_{1k}} z^{-1} + \boldsymbol{d_{2k}} z^{-2}}, \quad k = 1, ..., N_B,$$

где c_k,, d_{1k} и d_{2k} — коэффициенты фильтра, зависящие от номера полосы барковских частот и от частоты дискретизации F_S. Кроме того, в вычислениях фильтрации

используется однобайтное представление данных там, где это допустимо с точки зрения сохранения приемлемой точности вычислений. При этом среднее отклонение рассчитанной таким образом энергии сигнала по барковским полосам от энергии, вычисленной с помощью преобразования Фурье, не превышает 8%. Точность распознавания за счет этого ухудшается примерно на 0.7%. Но в результате требующаяся для вычисления барковского спектра производительность ЦСП сокращается до 7.6 MIPS (в инструкциях 8-разрядного микроконтроллера).

Дополнительным резервом сокращения требующейся производительность ЦСП является упрощение процедуры квантования характеристических векторов. Использование в ней Евклидова расстояния для определения вектора из кодовой книги, ближайшего к квантуемому вектору параметров, требует применения операций умножения двухбайтных чисел. При реализации программы идентификации диктора на 8-разрядном микроконтроллере такая операция оказывается чрезмерно трудоемкой. Поэтому в качестве меры близости квантуемого вектора к табличному предлагается использовать манхеттенское расстояние:

$$D_M(k) = \sum_{i=1}^{m} \left| X_i - Y_{ki} \right|, \quad k = 1,\ldots,N_{CB},$$

где X_i и Y_{ki} – i-ые координаты квантуемого вектора и k-го вектора кодовой книги, m – размерность вектора параметров, N_{CB} – количество векторов в кодовой книге.

Эксперименты показывают, что замена Евклидова расстояния на манхеттенское при использовании двухбайтного представления квантуемого вектора и векторов кодовой книги на 0.4% ухудшает точность распознавания, в несколько раз сокращая вычислительные затраты на процедуру квантования при реализации распознавателя на 8-разрядном микроконтроллере.

В результате предлагаемые упрощения алгоритмов полосовой фильтрации и квантования характеристических векторов и использование однобайтного представления данных там, где это допустимо с точки зрения сохранения приемлемой точности вычислений, позволяют реализовать распознаватель речи на 8-разрядных микроконтроллерах с производительностью до 10 MIPS, но с несколько худшей точностью распознавания (92–93% в шумах с SNR до 15 дБ).

В случае использования типового 16-разрядного ЦСП для реализации распознавателя последовательности изолированных слов замена БПФ на банк полосовых фильтров при вычислении барковского спектра речевого сигнала позволяет сэкономить вычислительные ресурсы (производительность и память), освободив их для возможного мультизадачного или многоканального режима работы ЦСП.

Исследование алгоритмической базы распознавателя изолированных слов в исходном и модифицированном вариантах показало, что около 80% алгоритмов могут быть эффективно реализованы на параллельной архитектуре, примером которой является РРП. Это является следствием как свойства фундаментальной параллельности, присущей процессам восприятия и обработки звуковой информации [2], так и векторного характера многих выполняемых операций при преобразованиях и анализе речевого сигнала.

Дальнейший сравнительный анализ традиционных и модифицированных алгоритмов реализации распознавателя изолированных слов позволит выявить оптимальные решения для параллельной архитектуры РРП и аппаратной реализации процессорных элементов, составляющих основу этой архитектуры. Обе реализации распознавателя (полная, ориентированная на использование 16-разрядных ЦСП, и упрощенная, умещающаяся в 8-разрядный микроконтроллер низкой производительности) могут быть поддержаны архитектурой РРП. Для этого представляется целесообразным разрабатывать РРП в двух многопроцессорных вариантах: на базе элементарных 8-

разрядных ЦСП и на базе 16-разрядных ЦСП, поддерживающих одновременное исполнение двух 8-разрядных операций.

Заключение

Реализация традиционных алгоритмов распознавания изолированных слов на ЦСП обеспечивает высокую точность распознавания: 95% для словаря из 100 слов в условиях "белого" шума с SNR=15 дБ, соответствующем уровню уличного шума. Однако она предъявляет относительно высокие требования к производительности процессора (около 20 MIPS в терминах операций типового 16-разрядного ЦСП) и памяти данных на кристалле (более 6 кбайт). Замена БПФ банком полосовых фильтров позволяет существенно снизить требования к производительности ЦСП (до 12 MIPS) и его памяти данных (до 3 кбайт) при практически неизменном качестве распознавания.

Таким образом, задача распознавания речи может быть успешно решена с соблюдением современных требований к точности распознавания на основе относительно дешевых ЦСП, обладающих сравнительно небольшими производительностью и памятью данных. Снижение требований к точности распознавания до 92–93% делает принципиально возможным реализацию распознавателя на 8-разрядных микроконтроллерах низкой производительности (до 10 MIPS).

Разработанные алгоритмы распознавания изолированных слов позволят получить уточненные оценки эффективности реализации и использования отдельных структурных элементов архитектуры РРП, составить перечень команд РРП, требующих аппаратной реализации, выявить оптимальную степень параллельности архитектуры РРП.

Список литературы

1. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proc. of the IEEE, 1989, Vol. 77, No 2, pp. 257-286.

2. Рождественский Ю.В., Дьяченко Ю.Г. Оценка фундаментальной параллельности в системах обработки голосовых сигналов // Системы и средства информатики. Вып.12. – М.: Наука, 2002, с.250-254.

3. Исследование новой вычислительной парадигмы и разработка на ее основе логического проекта динамического многопоточного процессора обработки сигналов. // Отчет о НИР, ИПИ РАН, шифр темы "Сигнал", N Г.Р. 01.2.00 104927, 2001, 251с.

4. Hermansky H., and Morgan N. Rasta processing of speech" // IEEE Transactions on Speech and Audio Processing, Special issue on Robust Speech Recognition, 2(4): 578-589, Oct. 1994.

5. Hermansky H. Perceptual Linear Predictive (PLP) Analysis of Speech // J. Acoust. Soc. Am., 1990, pp.1738 - 1752.

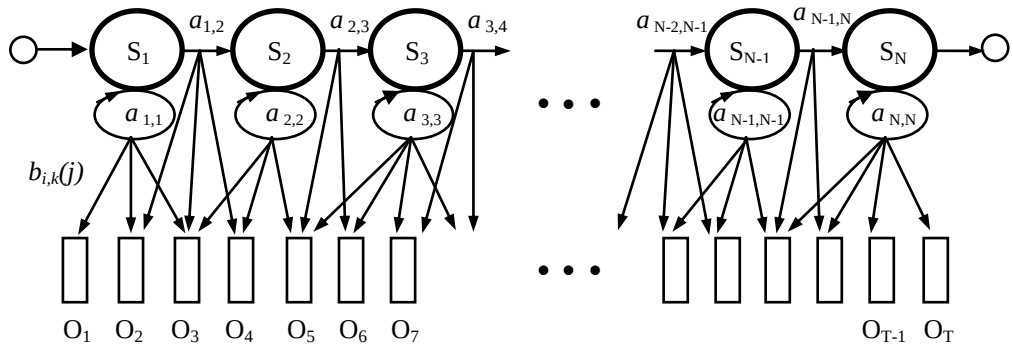

Рис.1. СММ слова

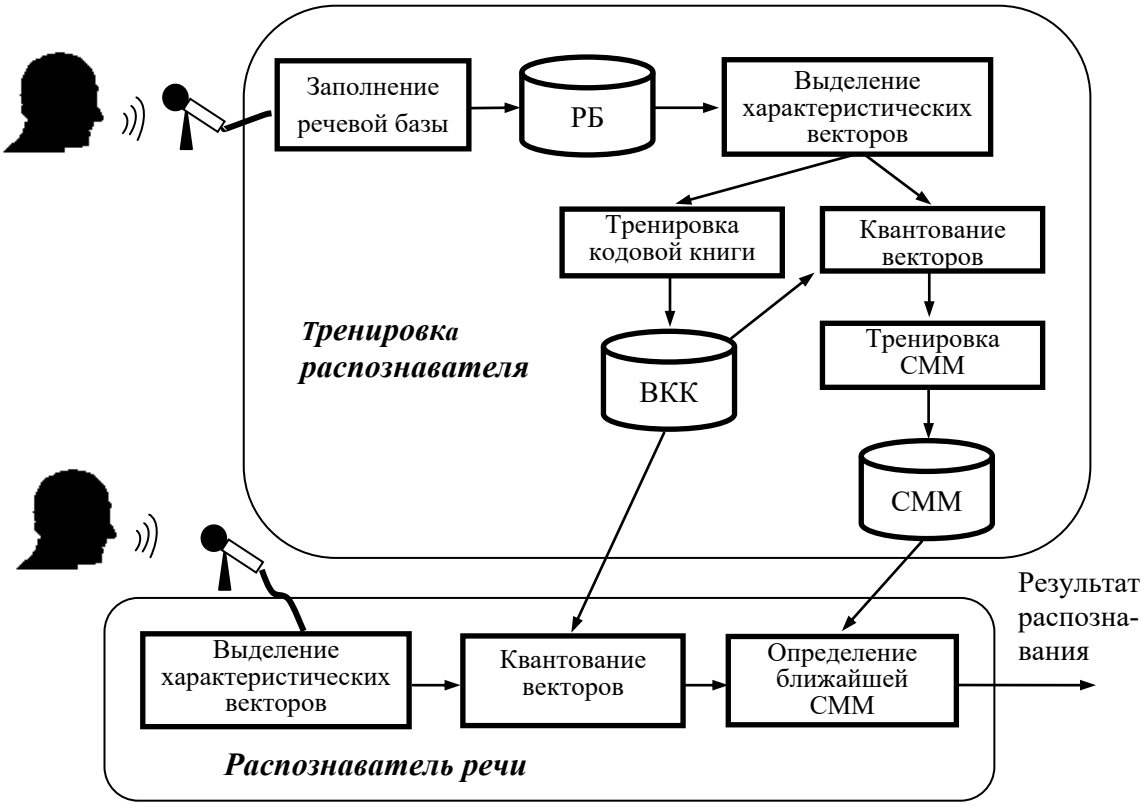

Рис.2. Структура системы распознавания речи

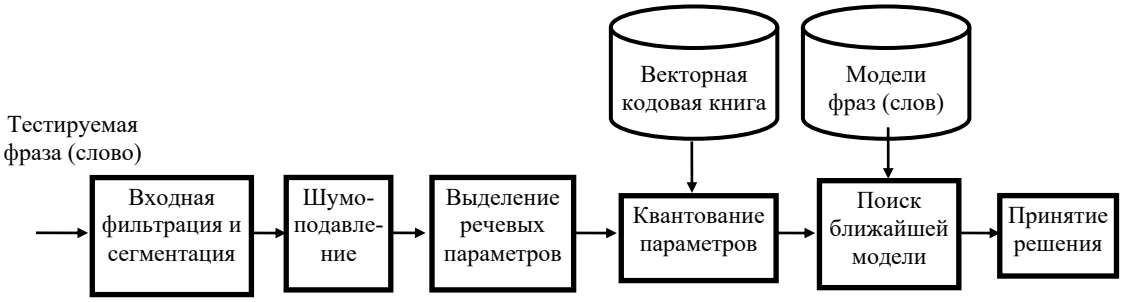

Рис.3. Блок-схема алгоритма распознавания речи