План выпускной квалификационной работы

1. Модели речевых сигналов
   1. Теория речеобразования
   2. Спектральные характеристики речевых сигналов (ОТ, форманты)
   3. Кепстральные характеристики

В соответствии с теорией речеобразования речь представляет собой акустическую волну, которая излучается системой органов: легкими, бронхами и трахеей, а затем преобразуется в голосовом тракте. Если предположить, что источники возбуждения и форма голосового тракта относительно независимы, речевой аппарат человека можно представить в виде совокупности генераторов тоновых сигналов и шумов, а также фильтров. Использование кепстрального анализа позволяет развернуть речевой сигнал, и получить информацию о состоянии артикуляционного аппарата, которая недоступна в частотной или временной области.

1. Обзор методов и алгоритмов распознавания речи
   1. Динамическое программирование

В случае реализации технологии распознавания с малым словарем (до 50 слов) применяют алгоритмы сравнения введенного образца с существующими в базе эталонами.

Одной из основных проблем распознавания речевых сигналов является тот факт, что одно и то же сочетание звуков, произнесенных несколько раз или же различными дикторами, может значительно отличаться по многим критериям: длительности, скорости произведения, форме огибающей, амплитуде и т.д. При разработке алгоритма распознавания с ограниченным словарем необходимо разработать такую решающую функцию, которая будет незначительно зависеть от этих критериев.

При этом наряду с разработкой базы эталонов, возникает проблема, связанная с тем, что одно и то же сочетание звуков может иметь различную длительность. Использование алгоритма динамического трансформирования времени позволяет избежать этой проблемы.

* 1. Скрытые модели Маркова

Метод скрытых марковских моделей. В качестве метода распознавания большинство современных систем используют метод скрытых марковских моделей [3, 4, 5]. Анализ применимости СММ для распознавания речи приводится в [6, 7]. Использование СММ для распознавания речи базируется на следующих предположениях: речь может быть разбита на сегменты (состояния), внутри которых речевой сигнал может рассматриваться как стационарный, переход между этими состояниями осуществляется мгновенно; вероятность символа наблюдения, порождаемого моделью, зависит только от текущего состояния модели и не зависит от предыдущих. Чаще всего используются СММ с тремя состояниями (рис. 1).



Рис. 1. СММ с тремя состояниями

СММ представляет собой конечный автомат, изменяющий свое состояние в каждый дискретный момент времени t. Переход из состояния si в состояние sj осуществляется случайным образом с вероятностью aij

В каждый дискретный момент времени модель порождает вектор наблюдений ot (который в конкретной задаче является вектором особенностей, полученным в преобразователе сигнала) с вероятностью bj (ot).

Распределение плотности вероятности наблюдений моделируется конечной гаусовской смесью с четырьмя компонентами. Каждая такая модель обозначает один из звуков русского языка или отсутствие звука (одна из моделей).

Алгоритмы распознавания ключевого слова используют эти модели для определения команд в потоке речи. Наиболее часто эта задача решается с помощью метода скользящего окна (sliding window) и метода моделей-заполнителей (filler models).

1. Исследование кепстральных коэффициентов речевых сигналов

Представление спектра сигнала в виде мел-частотных коэффициентов может успешно применяться в распознавании речи. Значения коэффициентов в шкале мел могут быть получены, анализируя значения коэффициентов в шкале Герц с последующим переходом при использовании выражения:

, (1)

где *f* – значение частоты в Герцах,

*В*(*f*) – значение частоты в мел, соответствующее частоте в Герцах *f*.

Тогда для оценки значений мел-частотных кепстральных коэффициентов на первом этапе необходимо оценить значения трансформанты Фурье анализируемого фрагмента сигнала вида:

, , (2)

где *xn* – анализируемый отрезок сигнала, длительностью *N* отсчетов,

*Nf* – количество точек Фурье.

При оценке логарифмов значений трансформант Фурье предлагается использовать треугольную оконную функцию вида:

 (3)

где *fm* – граничная частота *m*-го окна.

Окна предлагается располагать равномерно относительно шкалы мел, т.е. в шкале мел граничные частоты определяются с использованием выражения:

, , (4)

где *m* – номер треугольного окна,

*M –* число треугольных окон, равномерно расположенных в шкал мел,

*B*(*f*1) – нижнее значение частоты в шкале мел, рассчитанное с использованием выражения (1),

*B*(*f*в) – верхнее значение частоты в шкале мел, рассчитанное с использованием выражения (1).

Тогда в шкале Герц граничные частоты имеют вид:

, (5)

где *B*(*fm*) – граничные значения частоты в мел.

Для оценки кепстральных коэффициентов необходимо оценить значения логарифмов результата дискретного преобразования Фурье:

, , (6)

где *Xk* – значения трансформанты Фурье,

*M* – число треугольных окон, равномерно расположенных в шкале мел,

*Hm*,*k* – значения оконной функции вида (3).

Затем к полученным результатам применяются дискретное косинусное преобразование:

, , (7)

где *M* – количество треугольных окон, равномерно распределенных в шкале мел,

*Sm* – значение результата логарифмирования вида (6).

Исследование особенностей значений кепстральных коэффициентов для различных звуков русской речи показало, что распределение кепстральных коэффициентов зависит от типа звука. Таким образом, эта особенность может быть использована при разработке решающих функций распознавания речевых сигналов.

1. Исследование алгоритмов распознавания речи на основе динамического программирования.

Простейшим способом распознавания речевых сигналов при использовании ограниченного словаря видится сравнение анализируемого фрагмента сигнала с базой эталонов и принятие решения на основе наименьшего отклонения от какого-либо эталона. При этом наряду с разработкой базы эталонов, возникает проблема, связанная с тем, что одно и то же сочетание звуков может иметь различную длительность. Использование алгоритма динамического трансформирования времени позволяет избежать этой проблемы.

Алгоритм динамического трансформирования времени (DTW) вычисляет оптимальную последовательность трансформации (деформации) времени между двумя временными рядами. Алгоритм вычисляет оба значения деформации между двумя рядами и расстояние между ними.

При этом необходимо учитывать три условия, налагаемых на DTW алгоритм для обеспечения быстрой конвергенции:

• Монотонность – путь никогда не возвращается, то есть: оба индекса, *i* и *j*, которые используются в последовательности, никогда не уменьшаются.

• Непрерывность – последовательность продвигается постепенно: за один шаг индексы *i* и *j*, увеличиваются не более чем на 1.

• Предельность – последовательность начинается в (1,1) и заканчивается в (*I*,*J*).

В рамках данной работы при реализации алгоритма динамического трансформирования используются коэффициенты корреляции Пирсона, рассчитанные для кепстральных коэффициентов отрезков сигнала:

, , , (8)

где *с*1, *с*2 – кепстральные коэффициенты соответственно первого и второго сигнала,

*I*, *J* – длительности соответственно первого и второго сигнала,

*М* – количество кепстральных коэффициентов, используемых для анализа

Для оценки меры различий речевых сигналов может быть использовано остаточное после устранения временных различий расстояние между *А* и *В*:

, (9)

где *K* – количество операций накопления расстояний в матрице значений *gK*(*I*,*J*),

*gK*(*I*,*J*) – матрица расстояний вида:

. (10)

Для исследования данного алгоритма была составлена база эталонов из 10 слов (числительные от 0 до 9). В качестве исходного сигнала использовались записи речевого сигнала с частотой дискретизации 16кГц и разрядностью кода 16 бит. Исследование вероятностей ошибочного принятия решения осуществлялось на основе анализа 10 повторений этих же числительных тем же диктором (100 образцов). С помощью данного алгоритма производится сравнение анализируемого сигнала с сохраненными в памяти компьютера эталонами. В результате выбирается пара с минимальной дистанцией и делается вывод о соответствии сигнала слову из словаря. На рисунке 2 представлен результат оценки меры различий вида (10) для слова «четыре».

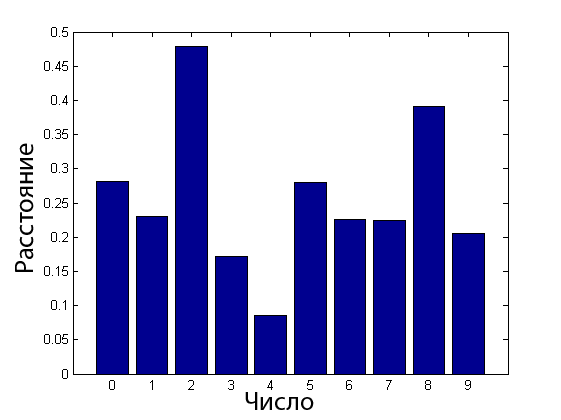


Рисунок 2 – Значения меры отличий для слова «четыре» (*N*=256, *M*=24)

Анализ рисунка показывает, что наименьшее значение отличия наблюдается для слова «четыре». Таким образом, в данном случае, алгоритм принимает верное решение.

В рамках данной работы были проведены исследования изменения вероятности ошибочного принятия решения при изменении количества кепстральных коэффициентов, используемых для анализа. Важно отметить, что при этом были рассчитаны все 24 коэффициента.

На рисунке 3 представлен график зависимости вероятности ошибочного принятия решения от количества анализируемых кепстральных коэффициентов.

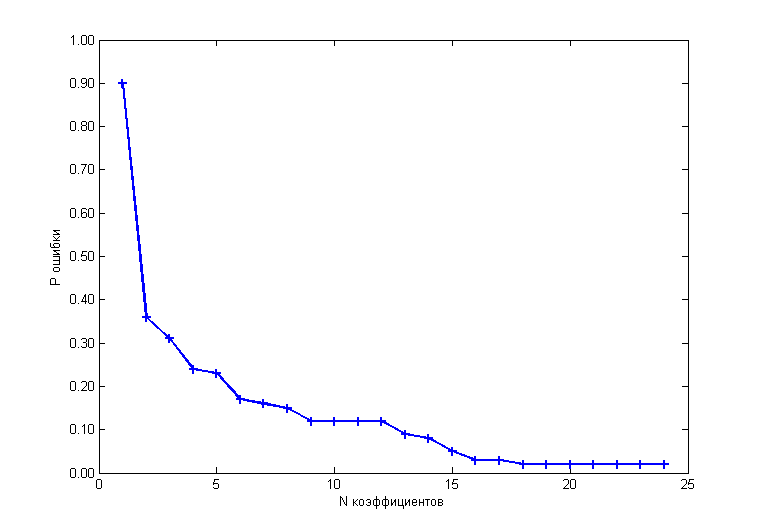


Рисунок 3 – График зависимости вероятности ошибочного принятия решения от количества анализируемых кепстральных коэффициентов (*N*=256)

Анализ результатов экспериментов показывает, что при анализе от 18 до 24 коэффициентов вероятность ошибочного принятия решения остается на одном и том же уровне. При этом увеличение количества анализируемых коэффициентов приводит к увеличению числа операций, выполняемых системой. Таким образом, при реализации данного алгоритма распознавания целесообразно использовать 18 кепстральных коэффициентов из 24 рассчитанных.