В настоящее время, с развитием компьютерных технологий, использование систем автоматического распознавания речи в качестве интерфейса приобретает все большую популярность. Однако, создание таких систем является нетривиальной задачей.

Проблеме распознавания речевых образов посвящено большое количество работ различных авторов. Успешное решение данной проблемы позволит осуществить частичную замену интеллектуальной деятельности человека действием автоматов.

Цели и задачи данной работы представлены на слайде

При распознавании речевых сигналов, как правило, оперируют не с исходным речевым сигналом, получаемым на выходе микрофона, а с так называемым описанием речевого сигнала, экономно представляющим речевой сигнал и содержащим информацию о том, что говорится.

Так, исходный речевой сигнал, который характеризуется объемом 256 кбит/с, как правило, описывается существенно меньшим объемом информации --- от 9600 до 600 и менее бит/с, однако все еще сохраняющим существенную информацию о том, что говорится, чтобы по ней отвечать на вопросы о распознаваемом классе.

Элементы  могут содержать компоненты, описываемые разнородными физическими величинами. Например, наряду с компонентами, представляющими форму амплитудного спектра речи или передаточную характеристику речевого тракта, могут быть компоненты, характеризующие интенсивность элемента, способ eгo образования (с участием голоса или только шума), относительную частоту основного тона и т. п. Последовательности , элементов  получают, анализируя речевой сигнал на интервале (окне) анализа продолжительностью  и перемещая это окно вдоль оси времени с шагом . Таким образом, интервалы анализа либо соприкасаются, либо перекрываются

Основой кепстрального анализа речевых сигналов является предположение, что речевой сигнал трактуется как сигнал на выходе линейной системы с медленно изменяющимися параметрами. Это предположение позволяет считать, что на коротких сегментах речевой сигнал можно рассматривать как сигнал на выходе линейной системы с постоянными параметрами, возбуждаемой либо последовательностью импульсов, либо случайным шумом. Проблема анализа сигнала сводится к измерению параметров модели и оценке изменения этих параметров с течением времени. Поскольку сигнал возбуждения и импульсная характеристика фильтра взаимодействуют через операцию свертки, задача анализа речи может рассматриваться как задача разделения компонент, участвующих в операции свертки. Такая задача иногда называется задачей обратной свертки

Если предположить, что источники возбуждения и форма голосового тракта относительно независимы, речевой аппарат человека возможно представить в виде совокупности генераторов тоновых сигналов и шумов, а также фильтров. Схема такой модели представлена на рисунке

Рассматриваемые фильтры имеют постоянные характеристики на временном интервале порядка 10-15 мс. Поэтому на каждом интервале фильтр можно характеризовать импульсной или частотной характеристикой или набором коэффициентов, если импульсная характеристика фильтра бесконечна. Такая модель позволяет применить для анализа речевых сигналов гомоморфную развертку.

Например, задан сигнал  на выходе фильтра. Требуется определить некоторую информацию о входном сигнале  и самом фильтре, например, о его импульсной характеристике .

Выходной сигнал определяется сверткой .

Т.к.  в частотной области, то прологарифмировав получаем выражение



Применив к нему обратное преобразование Фурье, можно получить выражение вида:



из которого методами линейной фильтрации может быть возможно выделить некоторые характеристики  и .

Также  может быть записано в виде:



Данное преобразование получило название <<кепстр>>. Аргумент q имеет размерность времени, но это особое, кепстральное время, поскольку  в любой момент q зависит от функции  исходного сигнала со спектром  заданной при . Иногда q называют «сачтота» или «кьюфренси» (анаграммы от русского <<частота>> или английского <<frequency>>).

Так как рассматриваемые в данной работе системы распознавания речи работают с дискретным представлением речевого сигнала, целесообразно привести запись кепстра в дискретной форме:

В литературе показано, что 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:



где  --- значение частоты в Герцах;

 --- значение частоты в мел, соответствующее частоте в Герцах .

На данном слайде представлена зависимость высоты звука в мелах от его частоты. Высота звука – величина, определяемая характеристиками человеческого слуха.

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

На данном слайде представлен пример графиков оконных функций для 12 окон анализа. Из графиков видно, что окна, равномерно расположенные в шкале мел, в частотной шкале расположены более плотно в области низких частот, что позволяет получить большее «разрешение» в низкочастотной области.

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



где  –- значения трансформанты Фурье,

 – число треугольных окон, равномерно расположенных в шкале мел,

 -– значения оконной функции

Затем к полученным результатам применяются дискретное косинусное преобразование:



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

 – значение результата логарифмирования

Исследование особенностей значений кепстральных коэффициентов для различных звуков русской речи показало, что распределение кепстральных коэффициентов зависит от типа звука.

Можно заметить, что для разных реализаций одного звука наборы коэффициентов схожи.

В проведенных экспериментах анализировались отрезки сигнала длиной в 256 отсчетов (16 мс при используемой частоте дискретизации 16 кГц) с перекрытием в 128 отсчетов, с целью обеспечения относительной стационарности анализируемого речевого отрезка. Количество точек Фурье устанавливалось равным длине отрезка.

Количество используемых коэффициентов установлено равным 24 на основании рекомендаций, приведенных в литературных источниках. Также было исследовано влияние количества используемых коэффициентов на точность распознавания речевых сигналов. В эксперименте рассчитывались все 24 коэффициента, а затем для описания сигнала в алгоритме распознавания использовались  коэффициентов.

На рисунке сплошной линией изображен график зависимости вероятности ошибки от количества коэффициентов, взятых с 1 по . Пунктирной линией -- с 24 по 

decrease\_legend.png Зависимость вероятности ошибки распознавания от количества используемых коэффициентов

Исходя из результатов эксперимента, можно предположить возможность использования меньшего количества коэффициентов в качестве описания сигнала, что позволяет уменьшить размер хранимого словаря и ускорить выполнение алгоритма распознавания.

Динамическое программирование в теории управления и теории вычислительных систем — способ решения сложных задач путём разбиения их на более простые подзадачи. Он применим к задачам с оптимальной подструктурой, выглядящим как набор перекрывающихся подзадач, сложность которых чуть меньше исходной.

В применении к задачам распознавания речи методы динамического программирования используются для определения степени схожести речевых сигналов. Как правило, подобное сравнение входного сигнала с имеющимся образцом имеет место в системах распознавания, работающих с ограниченным словарем (до 50 слов), но может также применяться на отдельных этапах принятия решений в составе комплексных систем

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.

Алгоритм динамического трансформирования времени (англ.

Dynamic Time Warp или DTW) вычисляет оптимальную

последовательность трансформации (деформации) времени

между двумя временными рядами. Алгоритм вычисляет оба

значения деформации между двумя рядами и расстояние между

ними.

Предположим, что есть две числовые последовательности  и . Длина двух последовательностей может быть различной.

Временные различия между A и B могут быть описаны с помощью некоторой последовательности :



где . Данная последовательность представляет собой функцию, которая позволяет отобразить временную ось A на временной оси B. Назовем ее функцией деформации.

На рисунке представлена функция F. По сути она напоминает классическую задачу динамического программирования о количестве редакторских правок (расстояние Левенштейна). Аргументом функции F является вектор c, содержащий индексы некоторого элемента матрицы d. Способ нахождения матрицы d представлен на следующем слайде

Алгоритм начинается с расчета локальных расстояний между

элементами двух последовательностей. В рамках данной работы

при реализации алгоритма динамического трансформирования

используются коэффициенты корреляции Пирсона,

рассчитанные для кепстральных коэффициентов отрезков

сигнала:

где с1, с2 – кепстральные коэффициенты соответственно

первого и второго сигнала;

I, J – длительности соответственно первого и второго сигнала;

М – количество кепстральных коэффициентов, используемых

для анализа.

В результате получаем матрицу расстояний, имеющую I строк и

J столбцов общих членов:

Минимальное остаточное расстояние между A и B, которое

остается после устранения временных различий, может служить

мерой различия речевых последовательностей A и B

Существует три условия, налагаемых на DTW алгоритм для

обеспечения быстрой конвергенции:

1) Монотонность – путь никогда не возвращается, то есть: оба

индекса, i и j, которые используются в последовательности,

никогда не уменьшаются.

2) Непрерывность – последовательность продвигается

постепенно: за один шаг индексы i и j, увеличиваются не

более чем на 1.

3) Предельность – последовательность начинается в (1,1) и

заканчивается в (I,J).

Также предлагается ввести дополнительное условие, названное <<окно выравнивания>>, определяющее дополнительные границы, в которых может лежать функция F

С учетом данных граничных условий, нахождение минимальной функции F является типичной задачей динамического программирования.

Вычислительная сложность реализованного алгоритма

составляет порядка O(IJ), возможно, сложность может быть

уменьшена оптимизацией алгоритма.

На данном слайде представлена блок-схема реализованного алгоритма

Алгоритм выполняет поиск кратчайшего пути в матрице коэффициентов корреляции, построенной с помощью выражения (), между элементами  и . В соответствии с принципом динамического программирования задача разбивается на элементарные операции, в данном случае для каждой точки , находящейся в окне выравнивания выбирается одно из трех возможных направлений перехода: ,  или , такое чтобы накопленная сумма значений  была минимальной. Полученное минимальное значение  записывается в выбранный для перехода элемент, таким образом происходит накопление суммы. То есть в элемент  будет записана сумма значений матрицы , лежащих в точках кратчайшего пути. Чтобы использовать эту сумму в качестве меры схожести двух наборов коэффициентов, необходимо произвести ее нормировку. Для этого  делится на длину полученного кратчайшего пути, то есть количество шагов.

На данном слайде можно видеть пример найденной функции F

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 16 кГц и разрядностью кода 16 бит. Исследование вероятностей ошибочного принятия решения осуществлялось на основе анализа 10 повторений этих же числительных тем же диктором (100 образцов). В результате, на 100 повторений было обнаружено 2 ошибки распознавания, что позволяет говорить о достаточной степени точности выбранного алгоритма.

К достоинствам выбранного алгоритма можно отнести достаточно высокую точность распонавания при работе с небольшим словарем, относительно небольшую сложность реализации и устойчивость к различиям в скорости произношения анализируемых слов. Однако, данный алгоритм обладает и рядом недостатков, таких как ограниченность размера используемого словаря, чувствительность к точности выделения слова из речевого потока и влиянию шумов.

Так диаграмма значений меры различий для произнесенного слова <<четыре>>, представленная на рисунке 10а, меняется при добавлении к сигналу аддитивного белого гауссовского шума с отношением сигнал-шум, равным 60 дБ, к виду, представленному на рисунке 10б. Из анализа диаграмм ясно, что принятие решения затруднено влиянием шума. Уменьшение отношения сигнал-шум до 40 дБ приводит диаграмму к виду, показанному на рисунке 10в, где можно видеть ошибочно принятое решение.

Был проведен экспериментальный анализ зависимости точности распознавания от значения отношения сигнал-шум. Для этого вся база из 100 записанных слов сравнивалась с имеющимися эталонами, при этом в каждой итерации из 100 сравнений к анализируемым речевым сигналам добавлялся аддитивный гауссовский белый шум различной мощности. Результаты эксперимента приведены на рисунке

Из рисунка видно, что даже при относительно высоких значениях отношения сигнал-шум, когда человек еще уверенно воспринимает произнесенные слова, используемый алгоритм показывает неприемлемо высокую вероятность ошибки распознавания.

В ходе выполнения данной работы были решены все поставленные задачи. Изучены существующие подходы к решению проблемы распознавания устной речи, выбран, реализован и исследован один из применяемых алгоритмов.

Модели речевых сигналов, использующиеся в системах распознавания речи, основываются на теории речеобразования и речевосприятия, поэтому основные теоретические понятия рассмотрены в первом разделе данной работы. Также в работе выполнен обзор основных методов, используемых для анализа сигналов в современных системах распознавания речи.

Из существующих алгоритмов распознавания устной речи для подробного анализа выбран один --- алгоритм динамического трансформирования времени (DTW), основанный на принципах динамического программирования. Выбор данного алгоритма объясняется возможностью его практической реализации в рамках выпускной квалификационной работы, а также его непосредственной применимостью к задаче распознавания речевых команд в системах с небольшим словарем (до 50 слов). Другие известные алгоритмы, использующие, например, скрытые модели Маркова или нейронные сети, обычно, являются частью более сложных систем, реализация которых требует существенных затрат.

Исследованный в работе алгоритм показал достаточно высокую точность (98%). При этом минимизировано влияние на точность распознавания громкости (за счет нормирования) и скорости произнесения (за счет методов динамического программирования). Наряду с достоинствами были выявлены и основные недостатки алгоритма, такие как низкая устойчивость к воздействию шумов. Вопросы дикторозависимости, которые не рассматривались в данной работе, могут быть предметом дальнейших исследований.