Актуальность работы

В настоящее время, с развитием компьютерных технологий, использование систем автоматического распознавания речи в качестве интерфейса приобретает все большую популярность. Создание таких систем является нетривиальной задачей. Успешное решение данной проблемы позволит осуществить частичную замену интеллектуальной деятельности человека действием автоматов.

Цели и задачи

Целью работы является реализация и исследование алгоритма распознавания речи.

Для достижения поставленной цели требуется решить следующие задачи:

1) Провести анализ моделей речевого сигнала;
2) Изучить подходы к решению задачи распознавания устной речи;
3) Реализовать алгоритм получения описания речевого сигнала и решающую функцию;
4) Исследовать реализованный алгоритм.

Описание речевых сигналов

При распознавании речевых сигналов, как правило, оперируют не с исходным речевым сигналом, получаемым на выходе микрофона, а с так называемым описанием речевого сигнала.

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

Элементы описания 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 его образования, относительную частоту основного тона и т. п.

Кепстральный анализ

Основой кепстрального анализа речевых сигналов является предположение, что речевой сигнал трактуется как сигнал на выходе линейной системы с медленно изменяющимися параметрами.

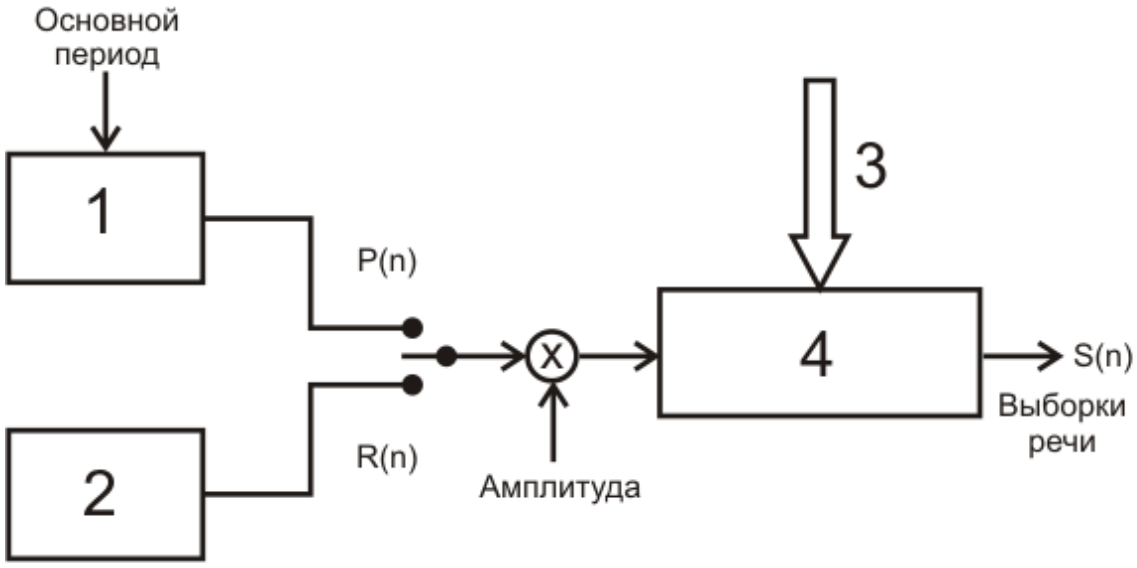


Рисунок 1 — Модель речевого аппарата в виде линейной системы – 1) Генератор импульсной последовательности; 2) Генератор случайных чисел; 3) Коэффициенты цифрового фильтра (параметры голосового тракта); 4) Нестационарный цифровой фильтр.

Кепстральный анализ

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.

Выходной сигнал определяется сверткой:

$$s_{\text{вых}}(t) = s_{\text{вх}}(t) \otimes h(t); \tag{1}$$

$$S_{\text{вых}}(\omega) = S_{\text{вх}}(\omega)H(\omega) \tag{2}$$

$$ln[S^2_{\text{вых}}(\omega)] = ln[S^2_{\text{вх}}(\omega)] + ln[H^2(\omega)]. \tag{3}$$

Кепстральный анализ

Применив к (3) обратное преобразование Фурье, можно получить выражение вида:

$$C(q) = C_s(q) + C_h(q) \tag{4}$$

из которого методами линейной фильтрации может быть возможно выделить некоторые характеристики $s_{\text{вх}}(t)$ и $h(t)$. Также $C(q)$ может быть записано в виде:

$$C(q) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \ln[S(\omega)]^2 e^{i\omega q}\, d\omega. \tag{5}$$

Данное преобразование получило название «кепстр».

Кепстральный анализ

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

$$C(n) = \frac{1}{N} \sum_{k=0}^{N-1} \ln |X(k)|^2 e^{i \frac{2\pi}{N} kn}, \ 0 \leq n \leq N-1. \qquad (6)$$

Мел-частотные кепстральные коэффициенты

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:

$$B(f) = 1125 \ln(1 + f/700) \tag{7}$$

где f — значение частоты в Герцах;
$B(f)$ — значение частоты в мел, соответствующее частоте в Герцах f.

Мел-частотные кепстральные коэффициенты

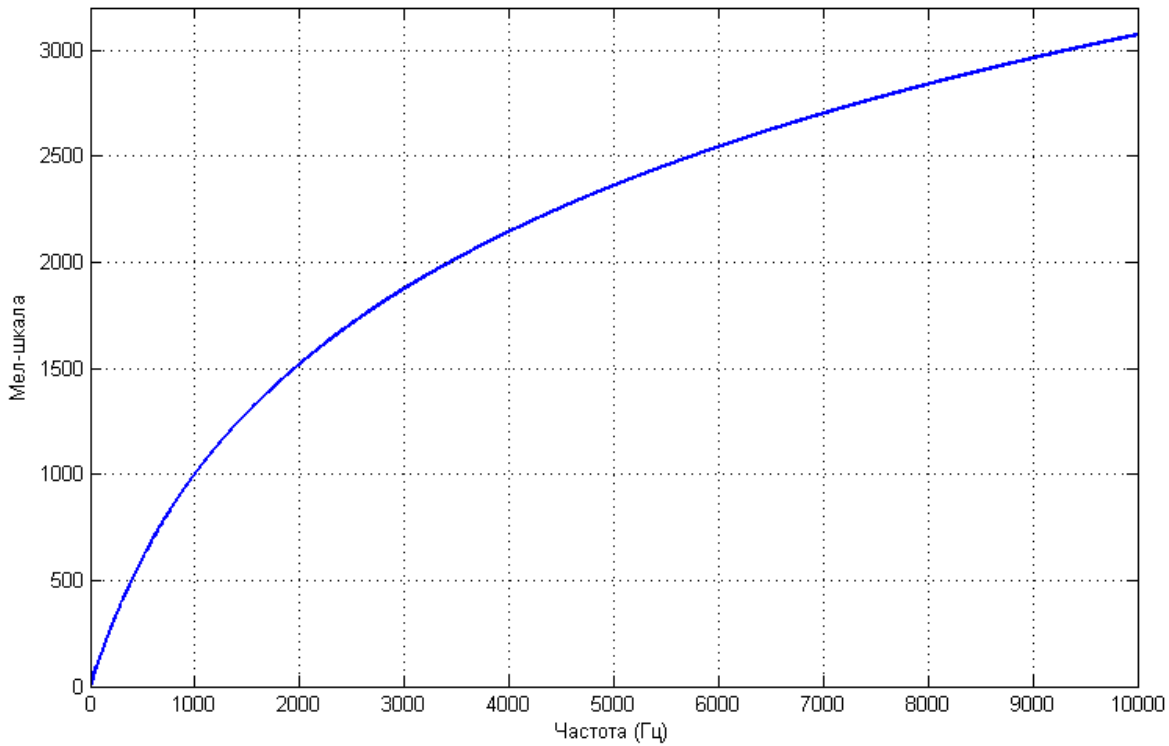


Рисунок 2 — Зависимость высоты звука в мелах от частоты

Мел-частотные кепстральные коэффициенты

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

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi i}{N} kn}, \quad 0 \leq k < N_f, \tag{8}$$

где x_n — анализируемый отрезок сигнала, длительностью N отсчетов; N_f — количество точек Фурье.

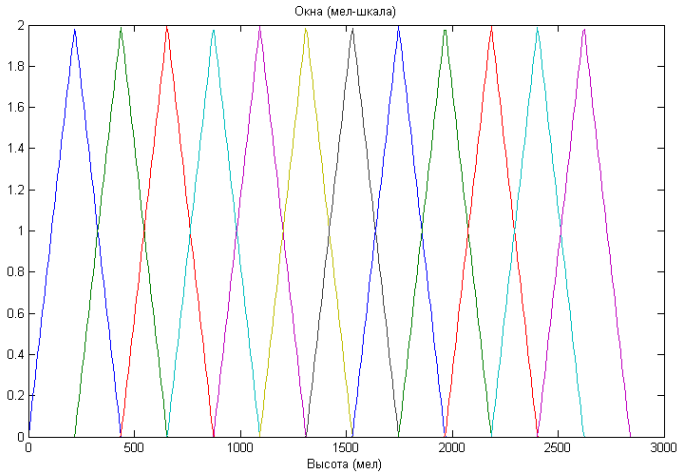
Мел-частотные кепстральные коэффициенты

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

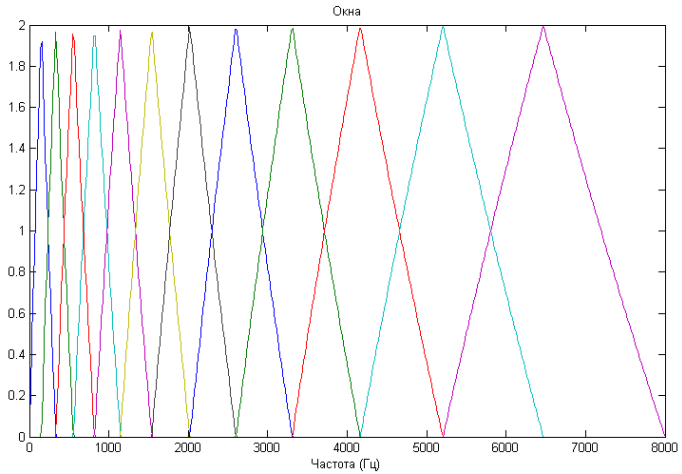
$$H_m = \begin{cases} 0, & k < f_{m-1} \\ \frac{(k - f_{m-1})}{(f_m - f_{m-1})}, & f_{m-1} \leq k < f_m \\ \frac{(f_{m+1} - k)}{(f_{m+1} - f[m])}, & f_m \leq k \leq f_{m+1} \\ 0, & k > f_{m+1} \end{cases} \tag{9}$$

где f_m – граничная частота m-го окна.

Мел-частотные кепстральные коэффициенты



(а)



(б)

Рисунок 3 — Графики оконных функций для 12 окон:
а) в шкале мел;
б) в шкале Гц

Мел-частотные кепстральные коэффициенты

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

$$S_m = \ln\left(\sum_{k=0}^{N-1} |X_k|^2 H_{m,k}\right), \quad 0 \le m \le M, \tag{10}$$

где X_k – значения трансформанты Фурье;
M – число треугольных окон, равномерно расположенных в шкале мел;
$H_{m,k}$ – значения оконной функции вида (9).

Мел-частотные кепстральные коэффициенты

Затем к полученным результатам применяются дискретное косинусное преобразование:

$$c_n = \sum_{m=0}^{M-1} S_m cos(\pi n(m + \frac{1}{2})/M), \quad 0 \leq n \leq M, \tag{11}$$

где M – количество треугольных окон, равномерно распределенных в шкале мел;
S_m – значение результата логарифмирования вида (10).

Мел-частотные кепстральные коэффициенты

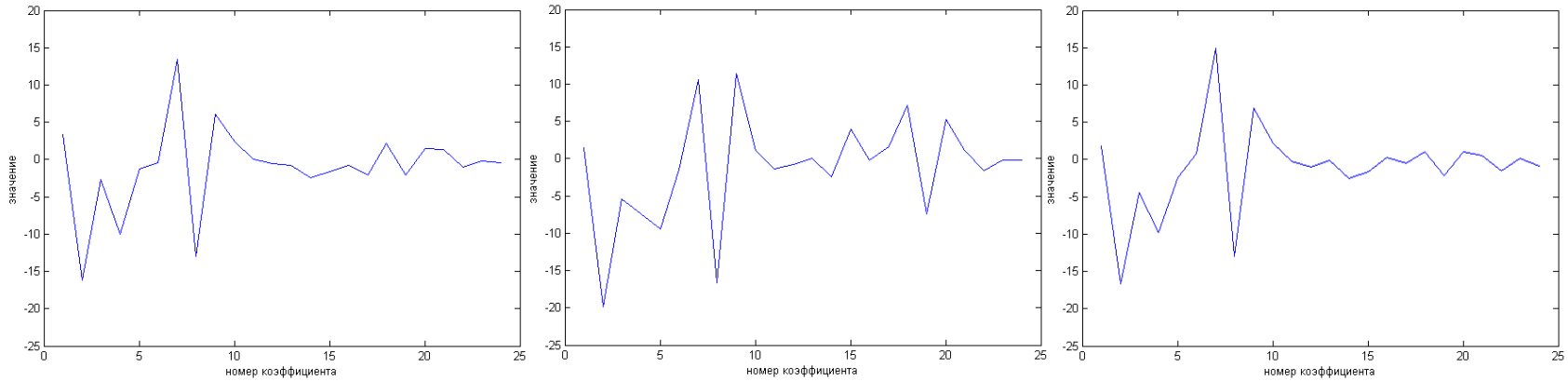


Рисунок 4 — Наборы мел-кепстральных коэффициентов для трех реализаций звука «а»

Мел-частотные кепстральные коэффициенты

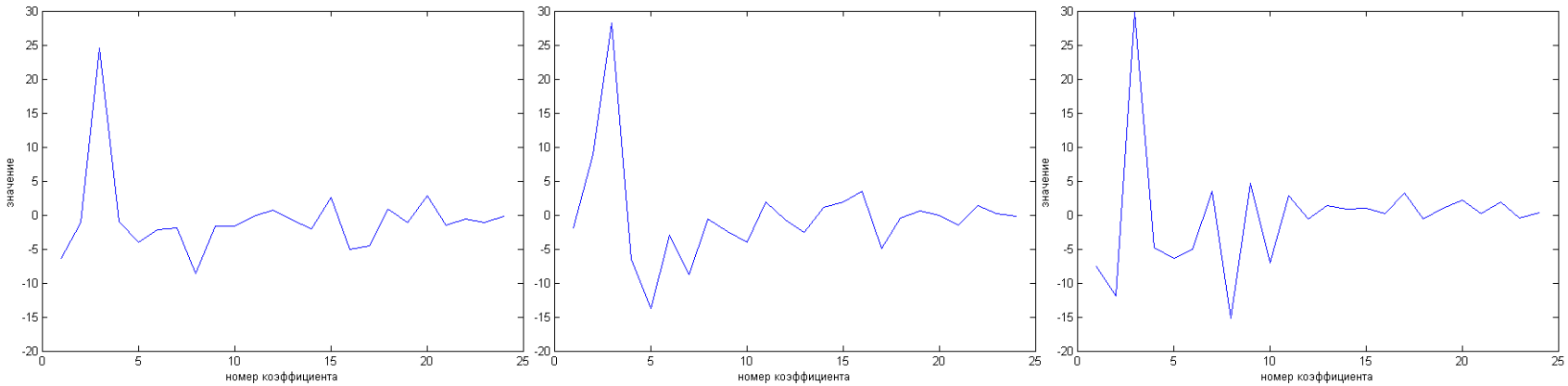


Рисунок 5 — Наборы мел-кепстральных коэффициентов для трех реализаций звука «и»

Мел-частотные кепстральные коэффициенты

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 литературе. 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 N_k коэффициентов.

Мел-частотные кепстральные коэффициенты

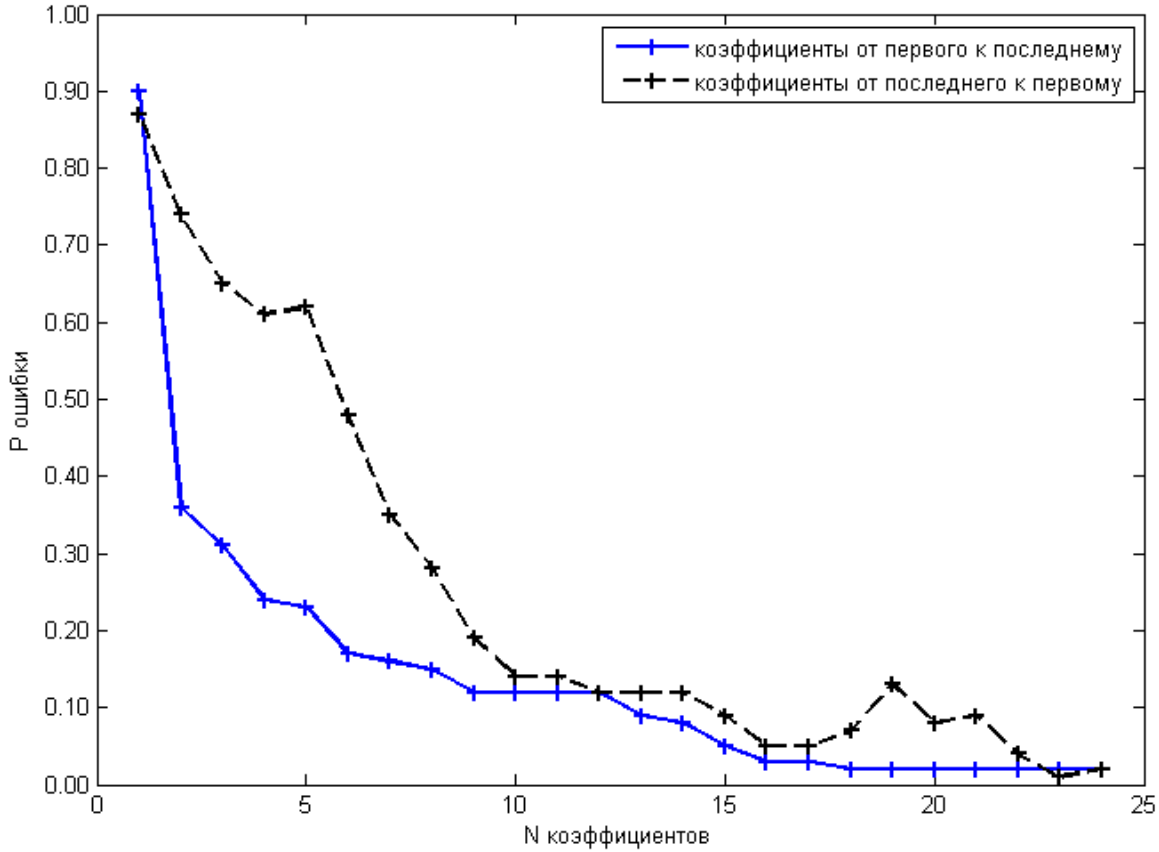


Рисунок 6 — Зависимость вероятности ошибки распознавания от количества используемых коэффициентов

Динамическое программирование

Динамическое программирование — способ решения сложных задач путём разбиения их на более простые подзадачи.

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.

Алгоритм динамического трансформирования времени

Одно и то же сочетание звуков может иметь различную длительность, что создает проблему для системы распознавания. Использование алгоритма динамического трансформирования времени позволяет избежать этой проблемы.

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.

Алгоритм динамического трансформирования времени

Предположим, что есть две числовые последовательности $A = a_1, a_2, \ldots, a_I$ и $B = b_1, b_2, \ldots, b_J$. Длина двух последовательностей может быть различной.
Временные различия между A и B могут быть описаны с помощью некоторой последовательности $c = (i, j)$:

$$F = c(1), c(2) \ldots, c(k), \ldots, c(K) \tag{12}$$

где $c(k) = (i(k), j(k))$. Данная последовательность представляет собой функцию, которая позволяет отобразить временную ось A на временной оси B. Назовем ее функцией деформации.

Алгоритм динамического трансформирования времени

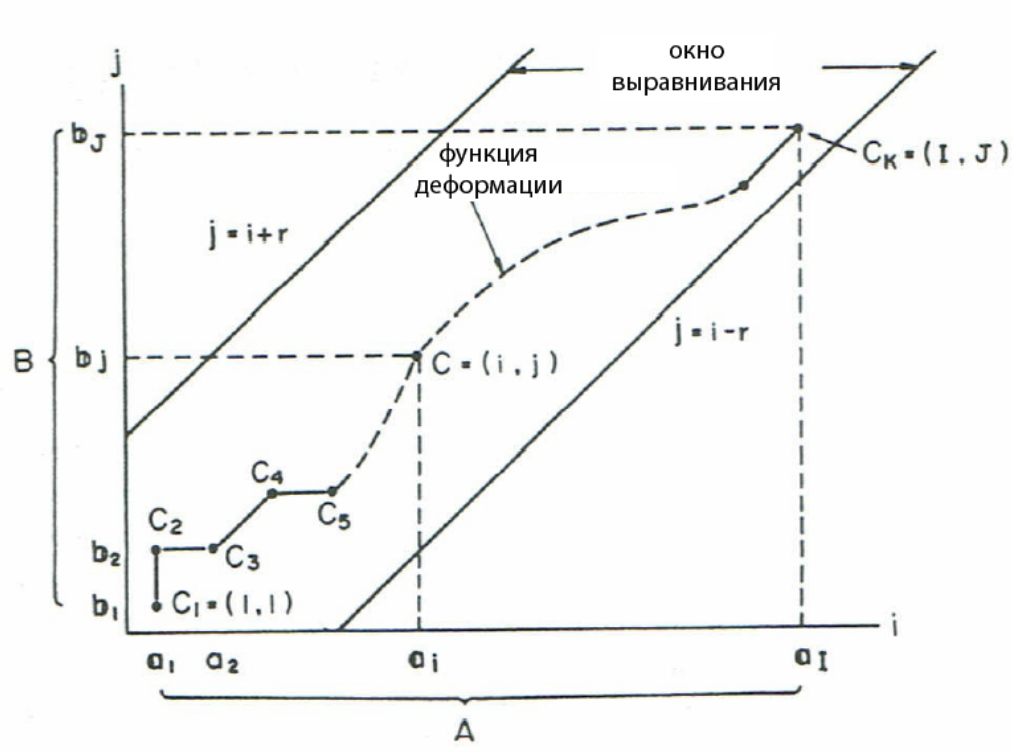


Рисунок 7 — Функция деформации и окно выравнивания

Алгоритм динамического трансформирования времени

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:

$$d(i,j) = 1 - \frac{\sum_{k-1}^{M} c1_{i,k} \cdot c2_{i,k}}{\sqrt{\sum_{k-1}^{M} c1_{i,k}^2 \cdot c2_{i,k}^2}}, \quad 1 \le i \le I, \; 1 \le j \le J, \qquad (13)$$

где c1, c2 – кепстральные коэффициенты соответственно первого и второго сигнала;
I, J – длительности соответственно первого и второго сигнала;
M – количество кепстральных коэффициентов, используемых для анализа.

Алгоритм динамического трансформирования времени

В результате получаем матрицу расстояний, имеющую I строк и J столбцов общих членов:

$$d(c) = d(i,j) \tag{14}$$

Минимальное остаточное расстояние между A и B, которое остается после устранения временных различий, может служить мерой различия речевых последовательностей A и B

$$Dist(A,B) = min_F \left[\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)} \right] \tag{15}$$

где $w(k)$ - неотрицательный весовой коэффициент.

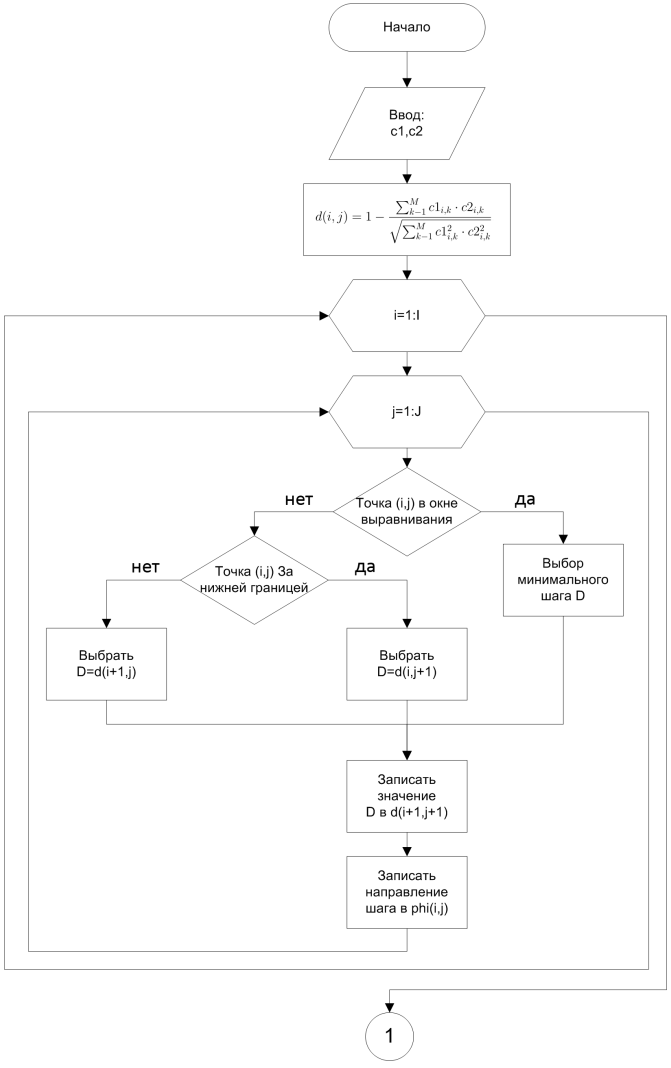
Алгоритм динамического трансформирования времени

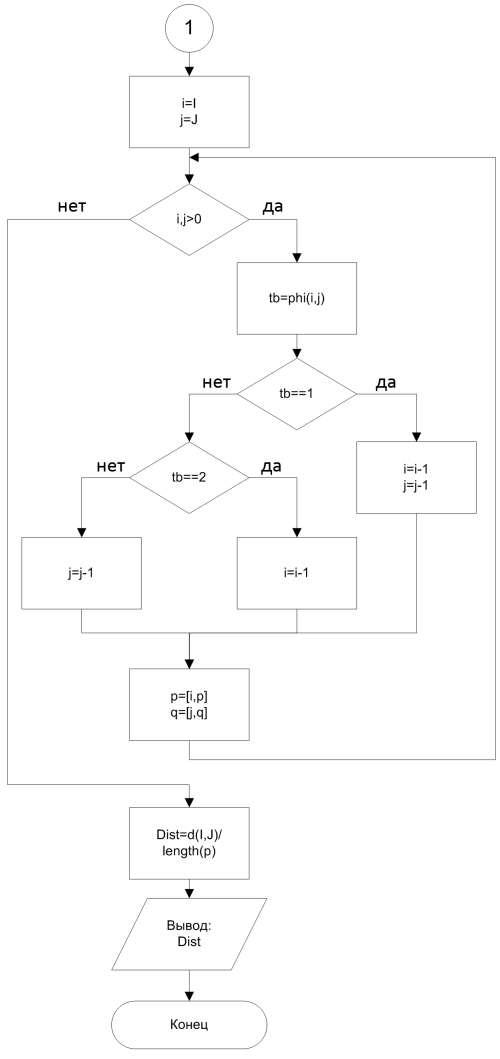
Существует три условия, налагаемых на DTW алгоритм для обеспечения быстрой конвергенции:

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).

Вычислительная сложность реализованного алгоритма составляет порядка $O(IJ)$, возможно, сложность может быть уменьшена оптимизацией алгоритма.

Блок-схема алгоритма





Алгоритм динамического трансформирования времени

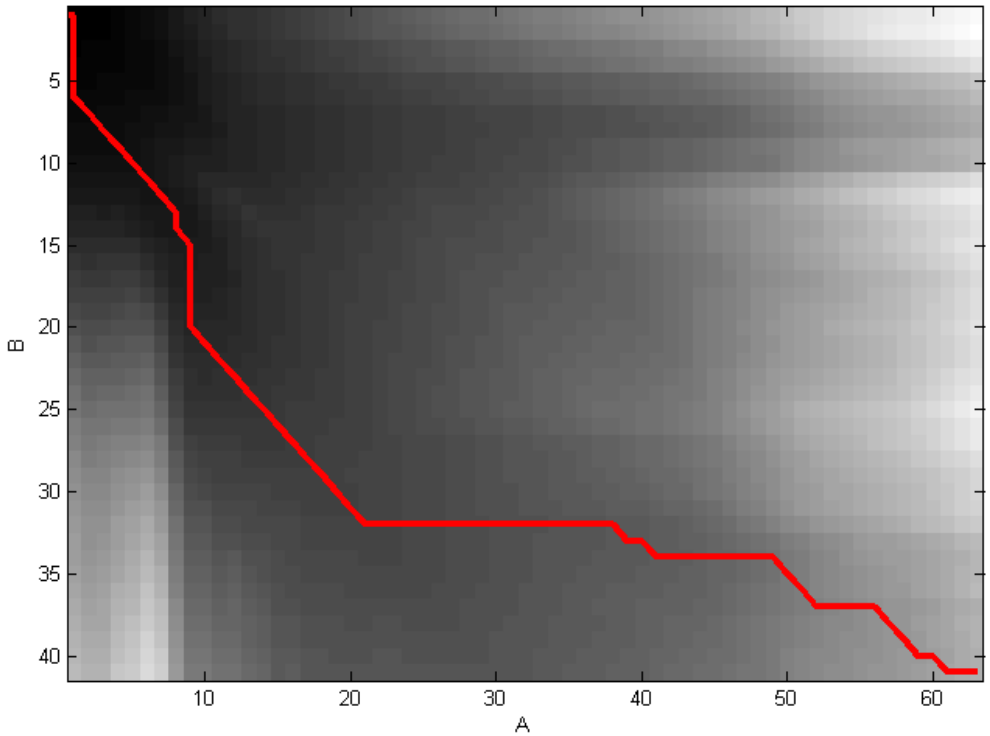


Рисунок 8 — Функция F вида (12), наложенная на матрицу d

Алгоритм динамического трансформирования времени

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.

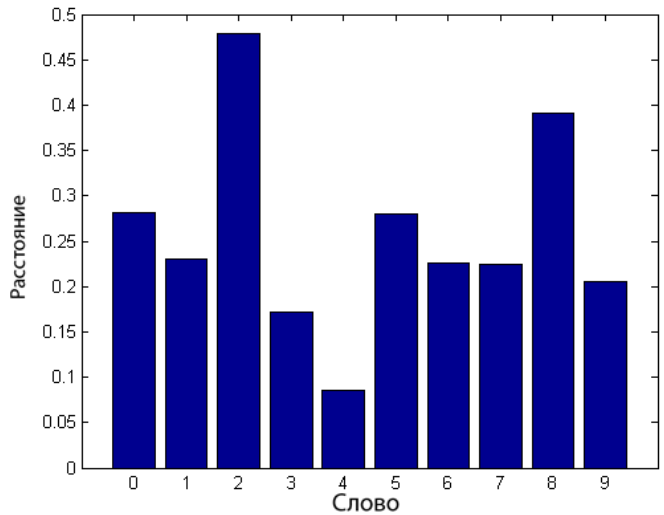


Рисунок 9 — Значения меры отличий для слова «четыре»

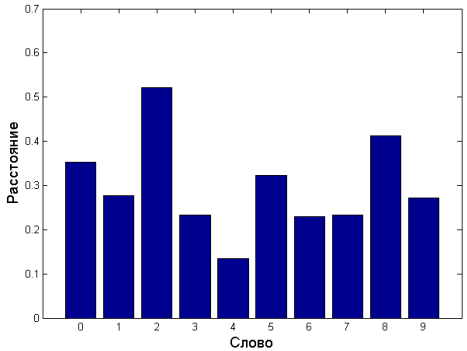
Алгоритм динамического трансформирования времени

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.

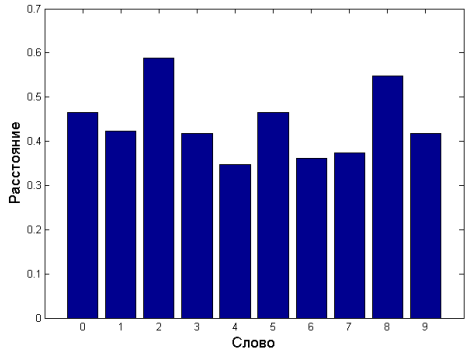
Алгоритм динамического трансформирования времени

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.

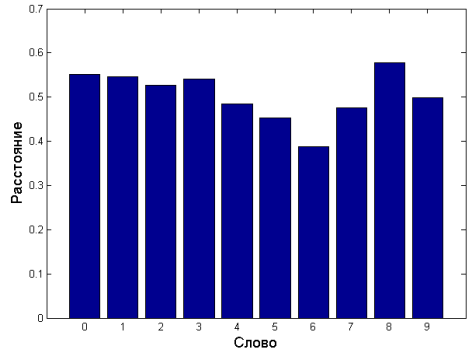
Алгоритм динамического трансформирования времени



(а) Значения меры отличий для слова «четыре» без влияния шумов



(б) Значения меры отличий для слова «четыре» (отношение сигнал-шум 60 дБ)



(в) Значения меры отличий для слова «четыре» (отношение сигнал-шум 40 дБ)

Рисунок 10

Алгоритм динамического трансформирования времени

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.

Алгоритм динамического трансформирования времени

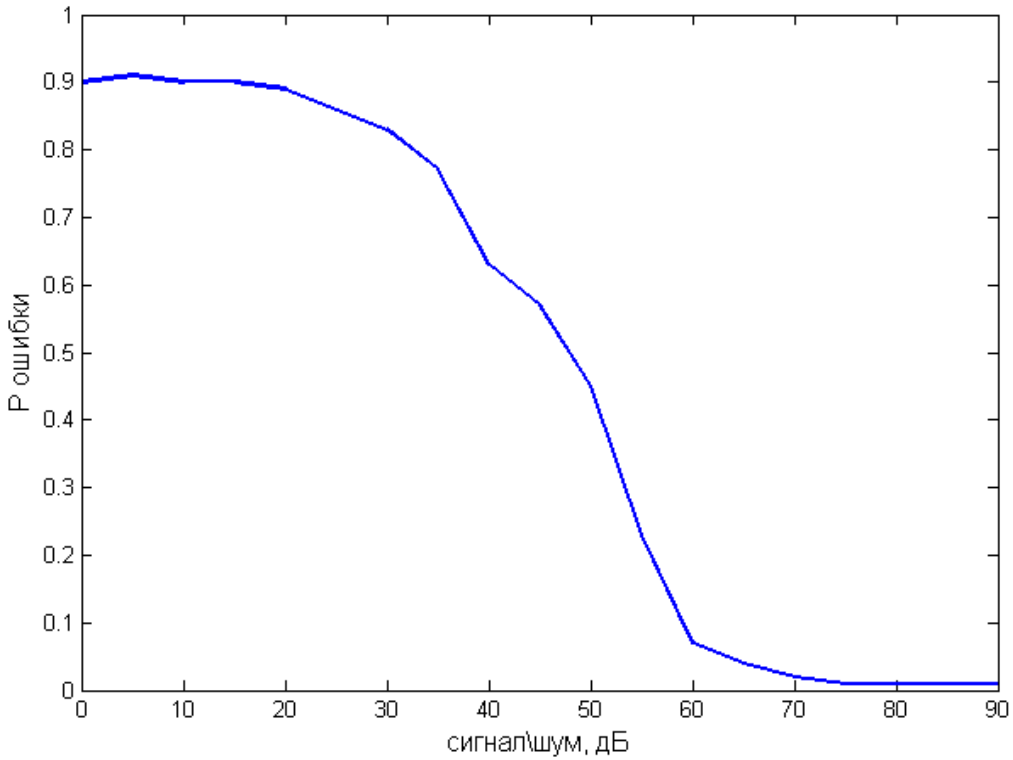


Рисунок 11 — Зависимость вероятности ошибки распознавания от отношения сигнал-шум

Заключение

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. 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. Другие известные алгоритмы, использующие, например, скрытые модели Маркова или нейронные сети, обычно, являются частью систем, реализация которых требует существенных затрат.