

Федеральное государственное бюджетное образовательное учреждение высшего
образования
«Московский авиационный институт (национальный исследовательский университет)»

Курсовая работа

Тема: «Информационно вычислительная система для
распознавания отдельно произнесенных слов»

Направление: 09.04.04.М1 «Программная инженерия»

Группа: 30-108М

Студент:

Кудрявцев А.А.

Научный руководитель:

Волков В. И.

Москва, 2017

Оглавление

Введение	3
Классификация систем распознавания речи	4
Распознавание отдельно произнесенных слов	6
Формирование словаря	6
Распознавание	7
Мел-кепстральные коэффициенты (MFCC)	8
Вычисление MFCC	8
Разложение в ряд Фурье.....	8
Расчёт mel-фильтров	8
Применение фильтров	11
Косинусное преобразование.....	11
Алгоритм динамической трансформации временной шкалы	12
Этапы алгоритма.....	12
Результаты экспериментов.....	14
Вывод.....	17

Введение

Задача распознавания речи состоит в восстановлении по звуковому сигналу слова естественного языка, произнесением которого является этот звуковой сигнал. Обычно она решается путем задания эталонов слов словаря и последующего сравнения звуковых сигналов с этими эталонами. Для решения задачи распознавания обычно сначала равномерно разбивают сигнал на окна одинаковой длины. Окна преобразуют из временной области в частотную (например, с помощью преобразования Фурье), чтобы близость окон относительно простых метрик соответствовала близости участков сигнала на слух. Затем решается задача нахождения соответствия между окнами звукового сигнала и окнами эталонов словаря. Сложность последней задачи заключается в том, что различные участки звукового сигнала в различных произнесениях одного и того же слова отличаются разной степенью сжатия или растяжения. Для решения задачи нахождения соответствия между окнами сигналов традиционно используются методы динамического программирования.

Предельные возможности компьютера по распознаванию речи связаны, прежде всего, с тем, что человек, которого можно взять за эталон распознающей системы, распознает осмысленную речь, а компьютеру в полной мере это не дано. ЭВМ не может с требуемой надежностью исправлять ошибки и неоднозначности распознавания, используя синтаксическую и семантическую связь слов предложения. Кроме того, человек использует зачастую дополнительную, незвуковую информацию. Ярким примером является чтение по губам, которое используют не только глухие люди, но и слышащие, для того чтобы лучше распознавать речь в шумной обстановке.

Дополнительно, картина осложняется тем, что все известные алгоритмы распознавания речи являются дикторозависимыми. После настройки на голос одного диктора распознающие системы дают удовлетворительные результаты распознавания для этого типа голоса, но хуже работают на других голосах. Надежность распознавания речи человеком, напротив, не зависит от типа голоса диктора.

Классификация систем распознавания речи

Согласно стандарту, принятому в области программирования систем распознавания речи, системы распознавания речи различают по следующим признакам:

Интервалы между словами

Если система распознает непрерывную речь, пользователь может произносить речевые фразы естественно, не делая паузы между словами. Непрерывное распознавание более предпочтительно, однако оно требует большей вычислительной мощности компьютеров.

Зависимость от диктора

Системы, обладающие относительной независимостью от диктора, позволяют пользователю работать с системой без предварительной настройки, однако улучшают надежность распознавания после обучения. Независимость от диктора таких систем обычно достигается за счет хранения звуковых эталонов для всех наиболее типичных голосов носителей данного языка. Это, безусловно, требует в несколько раз большей производительности и объема памяти. Настройка на голос диктора дикторонезависимых систем занимает обычно несколько часов. Это составляет главное неудобство для пользователя. Обычно дикторозависимые системы позволяют работать с относительной степенью надежности без предварительной настройки на голос конкретного пользователя. Третьей разновидностью систем по этому признаку являются системы, автоматически настраивающиеся на голос диктора по мере их использования. Системы последнего типа обладают следующими особенностями: им нужно знать, сделал ли пользователь ошибку, произнес конкретное слово, иначе обучение будет неверным; после настройки на одного диктора такие системы перестают надежно работать с другими голосами.

Степень детализации при задании эталонов

Различают алгоритмы, в которых в качестве эталонов используются целые слова, и алгоритмы, использующие эталоны элементов слов. Сравнение целых слов дает большую точность и скорость, однако требует значительно большего объема памяти. Алгоритмы сравнения элементов приходится применять в случае больших словарей, так как объем

требуемой памяти пропорционален количеству этих эталонных элементов слов и не зависит от объема словаря.

Размер словаря

Системы распознавания речи могут использовать большие и маленькие словари. Размер словаря системы распознавания почти не связан с реальным количеством слов, которое данная система может распознать. Он определяется количеством слов, требуемых для распознавания в данном конкретном состоянии системы. Системы работающие с маленькими словарями позволяет пользователю давать простые команды компьютеру. Для диктовки текстов необходимы большие словари (несколько десятков тысяч слов). Также системы диктовки могут учитывать контекст для определения активного подсловаря в конкретном состоянии, в данном случае система работает со словарем среднего размера.

Распознавание отдельно произнесенных слов

Система распознавания отдельно произнесенных слов принимает на вход аудиофайл и возвращает пользователю ответ. Ответом является наиболее близкое входному слову слово из словаря.

Формирование словаря

Чтобы добавить слово в словарь, пользователь должен отправить на сервер аудиофайл с этим словом и его текстовое представление. В базу данных (словарь) будет добавлен объект содержащий слово и массив значений, на основе которых будет вычисляться близость двух слов. В качестве таких значений было принято использовать мел-кепстральные коэффициенты — своеобразное представление энергии спектра сигнала.

Плюсы его использования заключаются в следующем:

- Используется спектр сигнала (то есть разложение по базису ортогональных синусоидальных функций), что позволяет учитывать волновую природу сигнала при дальнейшем анализе;
- Спектр проецируется на специальную mel-шкалу, позволяя выделить наиболее значимые для восприятия человеком частоты;
- Количество вычисляемых коэффициентов может быть ограничено любым значением (например, 12), что позволяет сжать фрейм и, как следствие, количество обрабатываемой информации;

Сигнал входного слова разбивается на небольшие отрезки (фреймы) и для каждого из них вычисляются мел-кепстральные коэффициенты.

Слова в словаре представляют собой объекты следующего вида:

id	Уникальный идентификатор
word	Слово
mfcc	Массив мел-кепстральных коэффициентов для каждого фрейма данного слова

Распознавание

Чтобы найти наиболее близкое входному слову слово из словаря необходимо найти массив мел-кепстральных коэффициентов для входного слова, а затем попарно сравнить его со значениями из словаря.

Очевидно, что длины сигналов сравниваемых слов будут отличаться, поэтому для сравнения массивов разной длины будем использовать алгоритм динамической трансформации временной шкалы. Результатом работы данного алгоритма является число – расстояние между двумя последовательностями, на его основе будет делаться вывод о близости произнесенного слова и слова из словаря.

Мел-кепстральные коэффициенты (MFCC)

Мел — психофизическая единица высоты звука, применяется главным образом в музыкальной акустике. Количественная оценка звука по высоте основана на статистической обработке большого числа данных о субъективном восприятии высоты звуковых тонов. Результаты исследований показывают, что высота звука связана главным образом с частотой колебаний, но зависит также от уровня громкости звука и его тембра. Звуковые колебания частотой 1000 Гц при эффективном звуковом давлении $2 \cdot 10^{-3}$ Па (то есть при уровне громкости 40 фон), воздействующие спереди на наблюдателя с нормальным слухом, вызывают у него восприятие высоты звука, оцениваемое по определению в 1000 мел. Звук частоты 20 Гц при уровне громкости 40 фон обладает по определению нулевой высотой (0 мел). Зависимость нелинейна, особенно при низких частотах (для «низких» звуков).

Вычисление MFCC

Пусть входной сигнал нормализован и разбит на небольшие временные промежутки – фреймы. Причем фреймы идут не строго друг за другом, а внахлест, т.е. соседние фреймы пересекаются между собой на 50%. Фреймы являются более подходящей единицей анализа данных, чем конкретные значения сигнала, так как анализировать волны намного удобнее на некотором промежутке, чем в конкретных точках.

Разложение в ряд Фурье

1. Рассчитаем спектр сигнала с помощью дискретного преобразования Фурье:

$$X[k] = \sum_{n=0}^{N-1} x[n] * e^{-2*\pi*i*k*n/N}, 0 \leq k < N$$

2. К полученным значениям применим оконную функцию Хэмминга, чтобы сгладить значения на границах фреймов.

$$H[k] = 0.54 - 0.46 * \cos(2 * \pi * k / (N - 1))$$

Таким образом, мы получим вектор следующего вида:

$$X[k] = X[k] * H[k], 0 \leq k < N$$

Расчёт mel-фильтров

Формула преобразования частоты в мел:

$$M = 1127 * \log\left(1 + \frac{F}{700}\right), \quad (1)$$

Обратное преобразование:

$$F = 700 * \left(e^{\frac{M}{1127}} - 1\right), \quad (2)$$

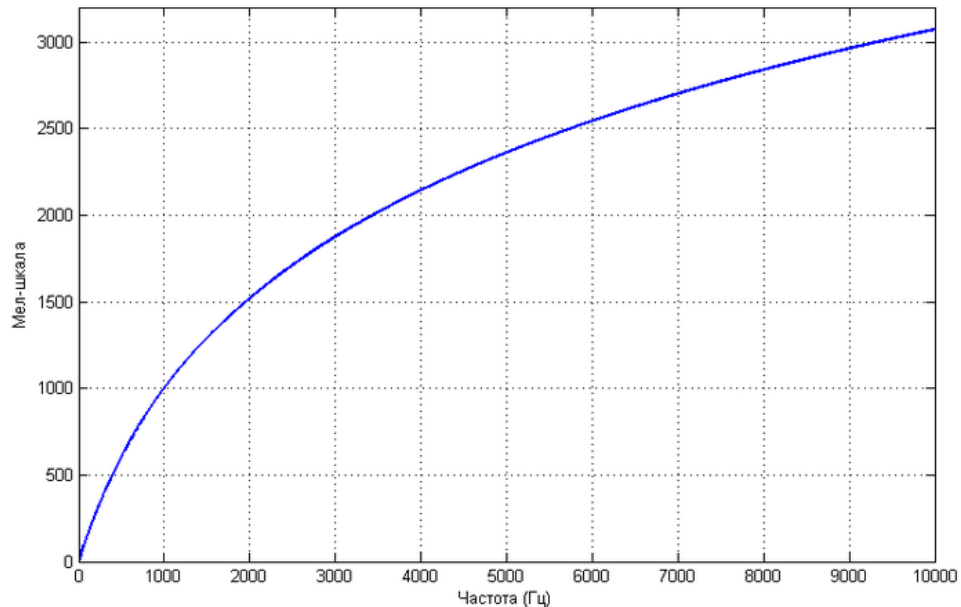


Рисунок 1. График зависимости мел/частота.

Допустим, у нас есть фрейм размером 256 элементов. Мы знаем, что частота звука в данной фрейме равна 16000hz. Предположим, что человеческая речь лежит в диапазоне [300; 8000]hz. Количество искоемых мел-коэффициентов положим $M = 10$.

Для того, что бы разложить полученный выше спектр по мел-шкале, нам потребуется создать M фильтров. По сути, каждый мел-фильтр это треугольная оконная функция, которая позволяет просуммировать количество энергии на определённом диапазоне частот и тем самым получить мел-коэффициент. Зная количество мел-коэффициентов и анализируемый диапазон частот мы можем построить следующий набор фильтров.

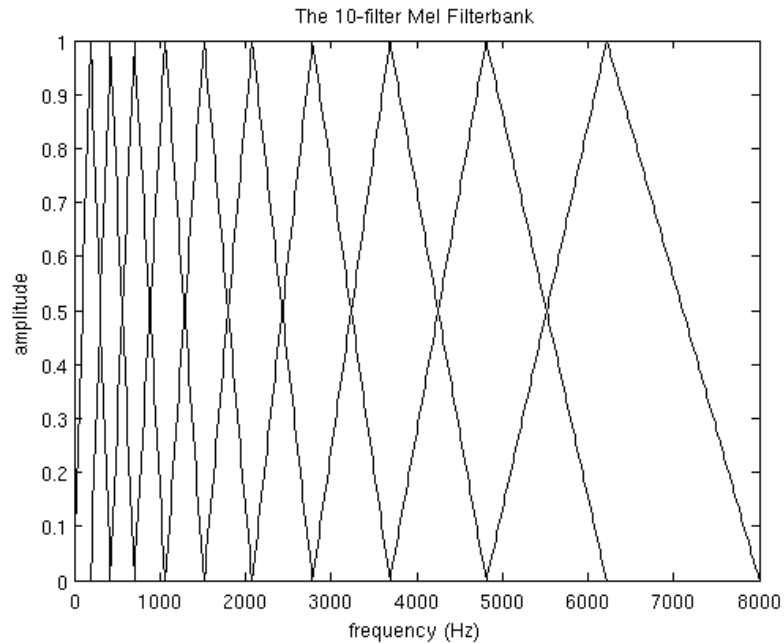


Рисунок 2. Mel-фильтры.

Пусть диапазон интересующих нас частот [300; 8000]. Согласно формуле (1) данный диапазон в мел-шкале эквивалентен [401.25; 2834.99]. Чтобы построить 10 мел-фильтров разделим данный диапазон на 10 равных частей, таким образом, получим 12 опорных точек:

$m[i] = [401.25, 622.50, 843.75, 1065.00, 1286.25, 1507.50, 1728.74, 1949.99, 2171.24, 2392.49, 2613.74, 2834.99]$

Используя формулу (2) переведем полученные точки обратно в герцы:

$h[i] = [300, 517.33, 781.90, 1103.97, 1496.04, 1973.32, 2554.33, 3261.62, 4122.63, 5170.76, 6446.70, 8000]$

Наложим полученную шкалу на спектр нашего фрейма.

$$f[i] = \text{floor} \left(\frac{(N + 1) * h[i]}{\text{rate}} \right),$$

где N – размер фрейма, rate - частота дискретизации.

В результате при $N = 256$ и $\text{rate} = 1600$ получим:

$f[i] = 4, 8, 12, 17, 23, 31, 40, 52, 66, 82, 103, 128$

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, f(m-1) \leq k \leq f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, f(m) \leq k \leq f(m+1) \\ 0, k > f(m+1) \end{cases}$$

Применение фильтров

Применение фильтра заключается в попарном перемножении его значений со значениями спектра. Результатом этой операции является mel-коэффициент. Поскольку фильтров у нас M , коэффициентов будет столько же. Однако нам нужно применить mel-фильтры не к значениям спектра, а к его энергии. После чего прологарифмировать полученные результаты. Считается, что таким образом понижается чувствительность коэффициентов к шумам.

$$S[m] = \log \left(\sum_{k=0}^{N-1} |X[k]|^2 * H_m[k] \right), 0 \leq m < M$$

Косинусное преобразование

Чтобы получить кепстральные коэффициенты, используется дискретное косинусное преобразование (DCT). Смысл DCT в том, что бы сжать полученные результаты, повысив значимость первых коэффициентов и уменьшив значимость последних.

$$C[l] = \sum_{m=0}^{M-1} S[m] * \cos \left(\pi * l * \frac{m + 0.5}{M} \right), 0 \leq l < M$$

Таким образом, получен набор из M mfcc-коэффициентов, которые могут быть использованы для дальнейшего анализа.

Алгоритм динамической трансформации временной шкалы

Алгоритм динамической трансформации временной шкалы (от англ. dynamic time warping) — алгоритм, позволяющий найти оптимальное соответствие между временными последовательностями. Впервые применен в распознавании речи, где использован для определения того, как два речевых сигнала представляют одну и ту же исходную произнесённую фразу. Впоследствии были найдены применения и в других областях.

Временные ряды — широко распространенный тип данных, встречающийся, фактически, в любой научной области, и сравнение двух последовательностей является стандартной задачей. Для вычисления отклонения бывает достаточно простого измерения расстояния между компонентами двух последовательностей (евклидово расстояние). Однако часто две последовательности имеют приблизительно одинаковые общие формы, но эти формы не выровнены по оси X. Чтобы определить подобие между такими последовательностями, мы должны «деформировать» ось времени одной (или обеих) последовательности, чтобы достигнуть лучшего выравнивания.

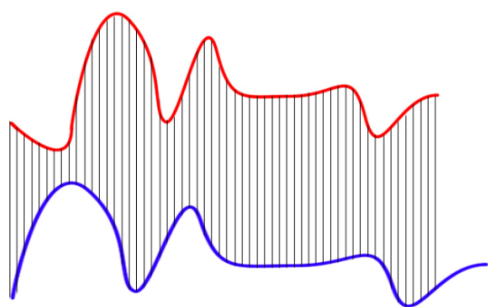


Рисунок 3. Евклидово расстояние

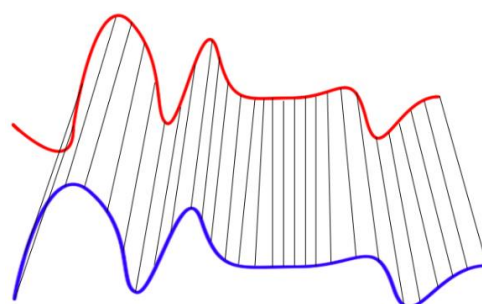


Рисунок 4. DTW

Этапы алгоритма

Рассмотрим два временных ряда Q длины n и C длины m:

$$Q = q_1, q_2, \dots, q_i, \dots, q_n;$$

$$C = c_1, c_2, \dots, c_j, \dots, c_m;$$

1. Построим матрицу расстояний d порядка $n \times m$, в которой элемент d_{ij} есть расстояние $d(q_i, c_j)$ между двумя точками q_i и c_j .
 - 1.1. В данном конкретном случае сравнения слов такими точками являются массивы кепстральных коэффициентов с фиксированной длиной M. Для

вычисления расстояния между ними используем евклидову метрику. Для точек $a = (a_1, \dots, a_M)$ и $b = (b_1, \dots, b_M)$ евклидово расстояние определяется формулой:

$$d(a, b) = \sqrt{\sum_{k=1}^M (a_k - b_k)^2}$$

2. Построим матрицу деформаций D , каждый элемент которой вычисляется исходя из следующего соотношения:

$$D_{ij} = d_{ij} + \min(D_{i-1, j}, D_{i-1, j-1}, D_{i, j-1})$$

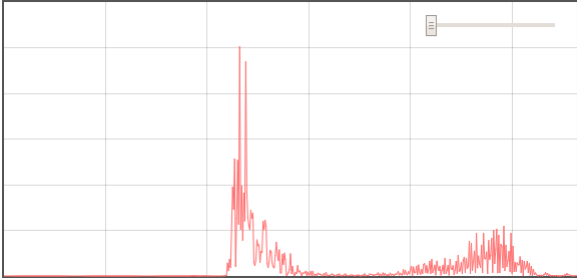
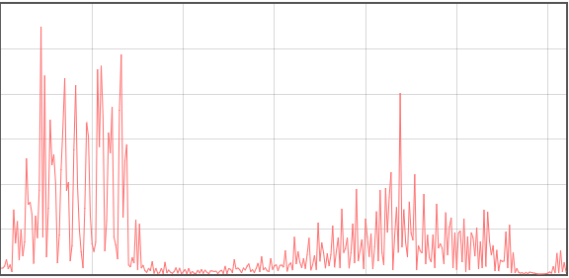
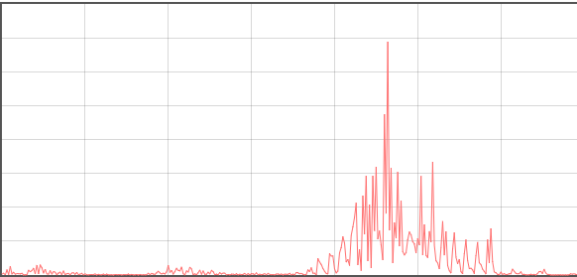
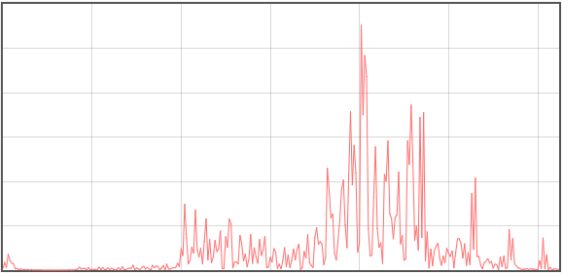
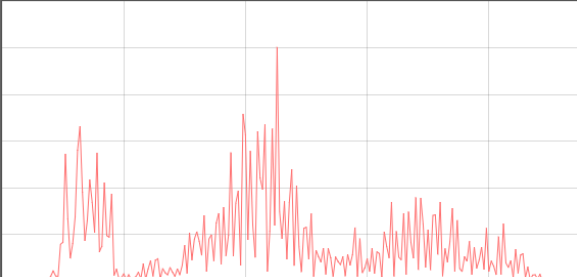
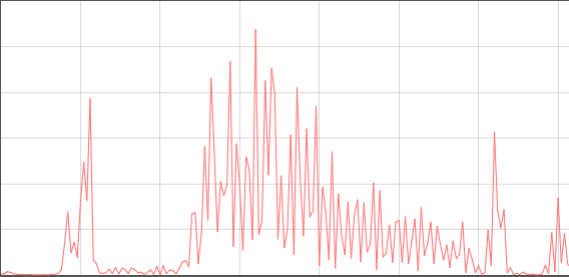
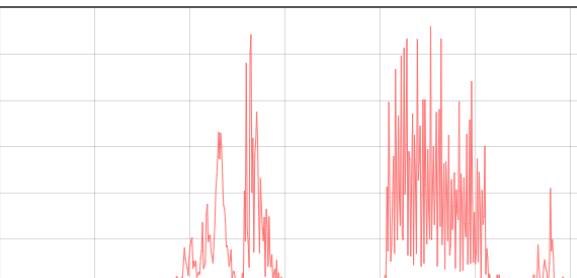
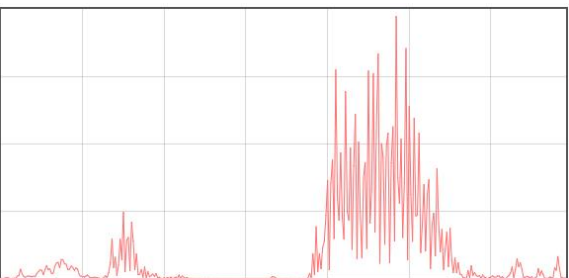
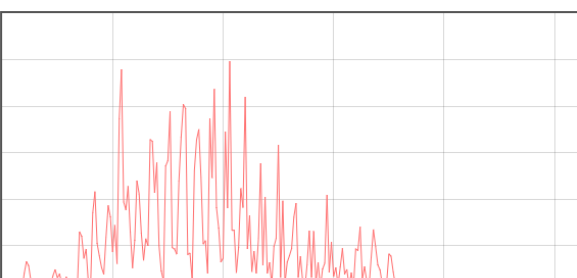
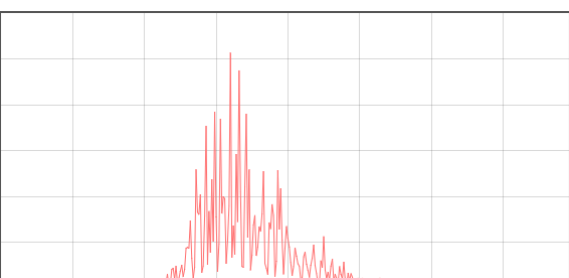
Расстояние между двумя последовательностями будет равно D_{nm} элементу матрицы расстояний.

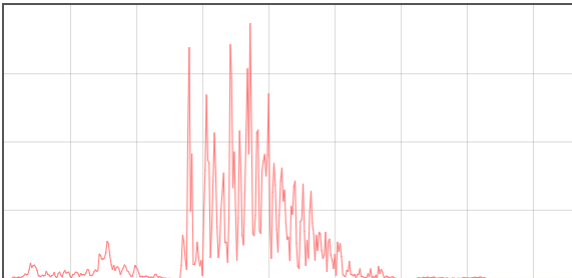
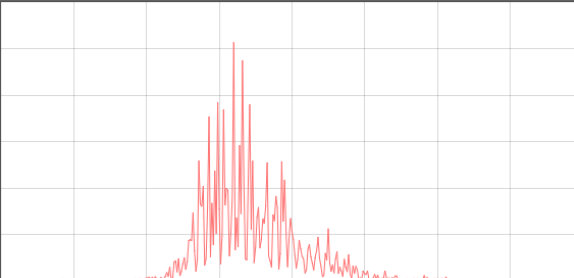
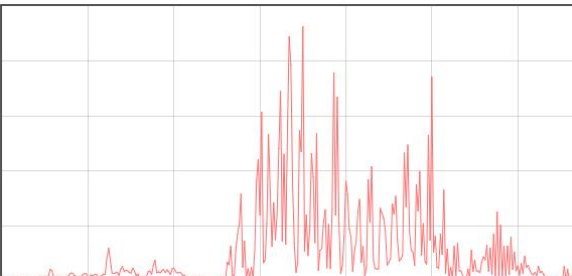
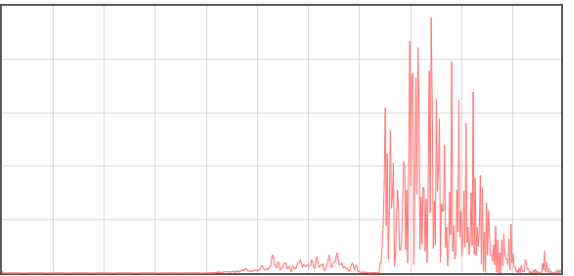
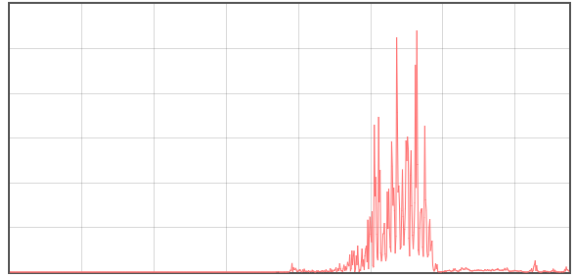
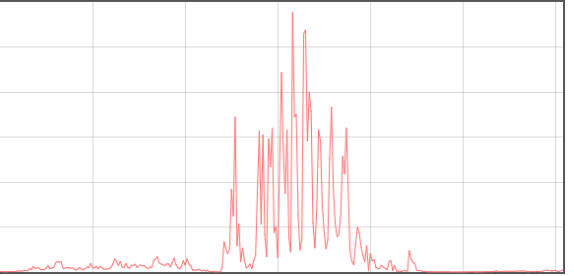
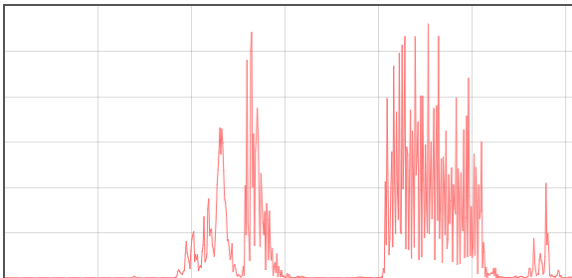
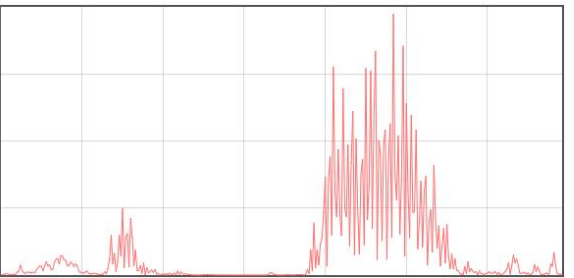
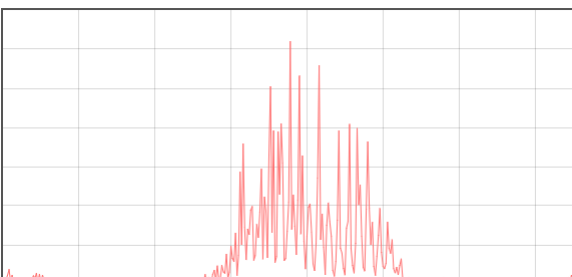
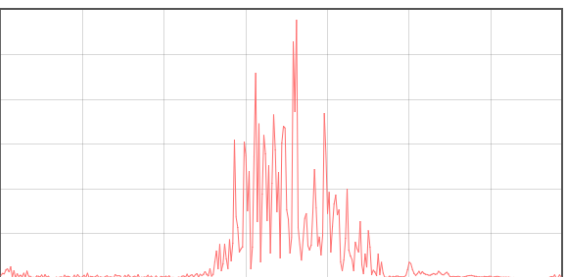
Условия, налагаемые на алгоритм динамической трансформации временной шкалы, для обеспечения быстрой конвергенции:

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 левом нижнем углу и заканчивается в правом верхнем.

Результаты экспериментов

Разработанная система была протестирована на словаре, состоящем из десяти слов («один», «два», «три», «четыре», «пять», «шесть», «семь», «восемь», «девять», «десять») и десяти отдельно произнесенных слов из словаря. В результате правильно было распознано 6 слов из 10. Предположительно улучшить данный результат можно увеличив размер словаря. Из таблицы ниже видно, что слова были распознаны не зависимо от того как была вычислена точка начала слова в аудиофайле, благодаря алгоритму динамической трансформации временной шкалы система нечувствительна к небольшим помехам не смотря на ошибки поиска границ слова. Ошибки же встречаются как правило при распознавании коротких слов из-за схожести их мел-частотных характеристик и ограниченности словаря.

✓	<p data-bbox="544 159 619 188">Один</p> 	<p data-bbox="1155 159 1230 188">Один</p> 
✓	<p data-bbox="552 497 611 526">Два</p> 	<p data-bbox="1163 497 1222 526">Два</p> 
✓	<p data-bbox="552 833 611 862">Три</p> 	<p data-bbox="1171 833 1230 862">Три</p> 
✓	<p data-bbox="528 1169 635 1198">Четыре</p> 	<p data-bbox="1139 1169 1246 1198">Четыре</p> 
✗	<p data-bbox="544 1505 619 1534">Пять</p> 	<p data-bbox="1139 1505 1246 1534">Девять</p> 

✕	<p data-bbox="539 152 625 185">Шесть</p> 	<p data-bbox="1145 152 1232 185">Девять</p> 
✓	<p data-bbox="547 499 617 533">Семь</p> 	<p data-bbox="1153 499 1224 533">Семь</p> 
✕	<p data-bbox="531 842 633 875">Восемь</p> 	<p data-bbox="1153 842 1224 875">Шесть</p> 
✕	<p data-bbox="531 1184 633 1218">Девять</p> 	<p data-bbox="1145 1184 1232 1218">Четыре</p> 
✓	<p data-bbox="531 1516 633 1550">Десять</p> 	<p data-bbox="1145 1516 1232 1550">Десять</p> 

Вывод

Предложенный метод распознавания отдельно произнесенных слов основан на преобразовании Фурье, вычислении мел-частотных кепстральных коэффициентов и вычислении расстояния с помощью алгоритма динамической трансформации временной шкалы. Данный набор технологий весьма распространен для решения задач распознавания и используется в некоторых известных системах распознавания речи. Мел-частотные кепстральные коэффициенты позволяют нам получить числовые значения, на основе которых мы можем сравнить два сигнала, выделяя из него лишь те частоты, которые наиболее важны при восприятии речи человеческим ухом, а алгоритм динамической трансформации временной шкалы позволяет нам сравнивать два сигнала разной длины. Результаты экспериментов проведенных в разработанной информационно-вычислительной системе показали, что она хорошо справляется со своей задачей, но также имеет и недостатки, над которыми в дальнейшем предстоит работать, среди них: примитивный способ определения границ слова, упрощенный алгоритм определения наиболее близкого слова из словаря, отсутствие в самой системе возможности записи звукового сигнала для распознавания.