Особенности задач распознавания звука

Л. П. Козлова
Санкт-Петербургский государственный
электротехнический университет
«ЛЭТИ» им. В.И.Ульянова (Ленина)
Санкт-Петербург, Россия
tigrenok59@mail.ru

О. А. Козлова
Санкт-Петербургский государственный университет
телекоммуникаций
им. проф. М.А. Бонч-Бруевича
Санкт-Петербург, Россия
k_olga_a@mail.ru

***Аннотация*. Среди задач распознавания образов последние годы стала выделяться задача технического слуха. Фактически системы распознавания речи уже стали частью не только моделирования различных процессов, но и внедрились в бытовые аспекты человеческой деятельности. Фактически принцип таких систем достаточно прост, однако процент ошибок работы неизменно высок, что означает сложности в применении на промышленном уровне. В статье будут рассмотрены особенности систем технического зрения, их сложности и проблемы.**

Ключевые слова: звук, распознавание образов, технический слух, сигнал

I. Введение

Среди прочих современных технических концепций, одной из наиболее часто используемых, является задача распознавания образов. Разделяясь на множество отдельных формализаций, она проникла не только в области науки и промышленности, но также используется и в простой человеческой жизни.

Фактически задача распознавания образов отвечает за выделение объекта из окружающей среды и присвоение его к определенному классу.

Наиболее часто эта задача применяется к распознаванию графических объектов. Например, можно столкнуться с подобными алгоритмами в задачах технического зрения.

Однако, последнее время не менее популярна задача выделения и распознавания звуковых сигналов. Реализации таких систем существуют уже давно, но практически применимы они стали лишь несколько лет назад. Это обусловлено высоким уровнем ошибок при работе алгоритмов.

II. Характеристики звука

Как было определено ранее, система распознавания предполагает в своей основе выделение примитива. Соответственно, для системы распознавания звука, необходимо понять, что такое звук, и как отличить один элемент от другого.

С точки зрения физики, звук это вибрация волны. Встречаясь с объектом, волна отражается от него, тем самым создавая изменения в воздушной среде.

Человек различает звук, когда изменения достигают уха и начинает воздействовать на барабанную перепонку. Далее мозг решает задачу распознавания образов.

Звуковая волна обладает рядом характеристик:

- форма;
- частота;
- амплитуда;
- фаза.

Базовой формой волны считается – синусоида. Именно она определяет остальные характеристики. Однако в реальности, как правило, волна, это не одна синусоида, а сочетание нескольких, что продемонстрировано на рис. 1.

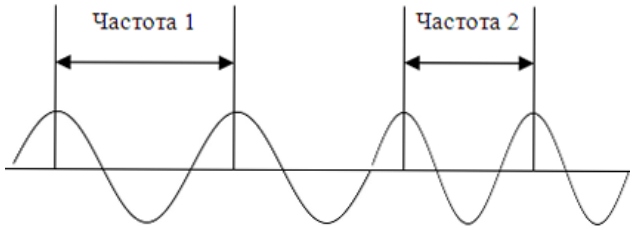

Рис. 1. Частота звуковой волны

Частота рассчитывает количество совершаемых колебаний в 1 секунду.

Амплитуда звуковой волны непосредственно связана с громкостью звука, а именно, чем больше амплитуда, тем громче звук.

Если две звуковые волны имеют одинаковую частоту и амплитуду, это означает, что они находятся в фазе. Изменение фазы происходит от 0 до 360. Тут 0 означает полное совмещение волн, а 360 – противофазу [1].

Зная амплитуду и частоту звуковой волны далее возможно превратить ее в набор математических значений, чтобы в дальнейшем проводить манипуляции именно с ними.

Тут следует обозначить основные проблемы, которые встречаются при работе со звуковыми файлами:

1. Возможное присутствие дополнительных звуковых волн. Такое может происходить из-за природного наличия дополнительных звуков (например, выделение пения конкретной птицы из общего фона лесных шумов) или ввиду различных звуковых эффектов (например, наличия эхо), т.п.
2. Один и тот же звук у разных, даже подобных объектов, может иметь отличия. Разная амплитуда приведет к тому, что один из звуков будет громче, второй – тише. Интенсивность частоты колебания звуковой волны повлияет на то, будет ли звук низким или высоким, что характерно, например, для человеческой речи.
3. Один и тот же звук может издаваться медленнее или быстрее. Например, когда человеку надо сказать что-то очень быстро, либо же речь ведется размеренно.
4. Качество аппаратной части, принимающей и передающей звуковой сигнал на дальнейшую обработку.
5. Если на одной записи происходит наложение двух сигналов, амплитуда которых не совпадает, то это приводит к полному исчезновению звука.

III. Задача распознавания звука

Обобщенный алгоритм системы распознавания звука можно записать следующим образом:

1. получение звукового сигнала;
2. устранение сторонних звуковых сигналов и усиление полезного;
3. идентификация полезного сигнала;
4. оценка качества распознавания.

По принципу построения весь массив алгоритмов делится на три большие группы, однако все они предполагают в своей основе эталонную базу, с элементами которой сравнивается сигнал. Причем сравнение может проводиться двумя способами: либо путем наложения спектрограммы сигнала, либо путем сравнения уже оцифрованных значений.

При работе со спектрограммами используется метод динамического временного деформирования (*Dynamic Time Warping – DTW*), при котором стартовый сигнал делится на несколько частей. Важной особенностью тут является то, что каждый следующий отрезок начинается не с того момента, где закончился предыдущий, а на некоторое количество фаз ранее, т.е. содержит в себе окончание предыдущего отрезка. Далее частоты сравниваются с помощью преобразования Фурье.

Действительно, пусть имеются две звуковые последовательности: $R = r_1, r_2, ... r_n$ и $P = p_1, p_2, ... p_m$, тогда для них можно рассчитать:

$$DTW(R,P) = \min\left\{\frac{\sum_{k=1}^{K} d(w_k)}{K}\right\},$$

где K – необходим для нормализации временных последовательностей разной длительности, $d(w_k)$ – матрица расстояний, рассчитывается по формуле:

$$d(w_k) = (r_i - c_j)^2 .$$

Сложность таких систем оценивается как $O(nm)$.

Другим методом, который применяется в системах распознавания звука, является использование скрытых марковских моделей (Hidden Markov Model – HMM).

Скрытая марковская модель использует систему конечных автоматов, которая состоит из некоторого количества состояний. Поскольку для данного алгоритма не важна их последовательность, такая особенность получила название скрытых состояний.

Отсюда в методе делаются два обязательных допущения:

1. Звуковой сигнал разбивается на фрагменты, каждому из которых, соответствует определенное состояние в модели, таким образом, что характеристики внутри одного фрагмента не меняются.

2. Состояние предыдущих и последующих фрагментов не влияют на вероятности друг друга. Вес имеет только текущее состояние.

Обозначим марковскую модель – λ и определим для нее следующие параметры: пусть матрица переходов между состояниями обозначается как $A = \{a_{ij}\}$, где a_{ij} – вероятность конкретного перехода между состояниями *i* и *j*; общее количество состояний в системе – *N*; *M* – значения, которые принимает система в любом из состояний; матрица определяющая вероятность выходных значений – $B = \{b_j(k)\}$, где $b_j(k)$ – определяет для состояния *j* вероятность выпадения параметра *k*; $\Pi = \{\pi_i\}$ – матрица определяющая вероятность попадания в начальное состояние, где π_i – вероятность попадание системы в состояние *i* в начальные момент времени. Тогда скрытая марковская модель определяется как $\lambda = \{A, B, \Pi\}$ [2].

Для того, чтобы при практическом применении система давала результаты, соответствующие критериям качества, необходимо решить также ряд подзадач, к которым отнесем:

- выбор алгоритма, позволяющего нормализовывать используемые вектора;
- определение параметров *N* и *M*;
- выбор алгоритма сегментации для начального этапа работы алгоритма;

- эталонная база должна быть обширна и содержать исчерпывающее количество примитивов и их сочетаний.

Следующим методом работы со звуковыми файлами является использование искусственных нейронных сетей.

Принцип действия системы показан на рис. 2.

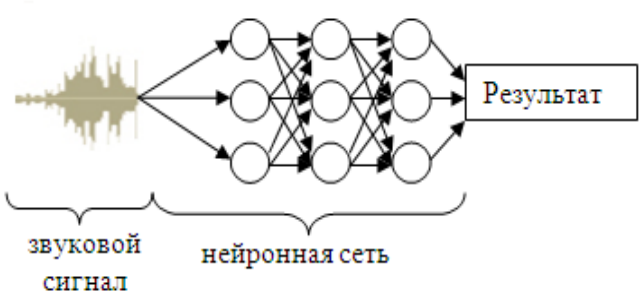

Рис. 2. Принцип распознавания звука на основе нейронной сети

На вход нейронной сети подается звуковой сигнал. Далее множество входных параметров подвергаются обработке, аналогичной классификации данных, после чего распознанная информация поступает на выход системы в соответствии с дальнейшими задачами.

Стоит заметить, что в сравнении с системами распознавания графических образов, системы технического слуха, основанные на нейронных сетях, еще отстают по качеству полученного результата. Это обусловлено, в первую очередь недостаточно обширной эталонной базой. Однако, учитывая приспособленность нейронных сетей к самообучению, а также возможность построение уникальной сети для каждой конкретной задачи, это одно из наиболее перспективных направлений для задач распознавания звуков [3].

IV. Заключение

Задачи распознавания слуха имеют обширный спектр применения в современном мире. В первую очередь такие алгоритмы используются для распознавания человеческой речи, например, для голосового управления различными системами.

Другим примером применения может служить выделение неких звуковых сигналов с общей записи.

В любом случае качество работы системы всегда зависит от объема и содержания эталонной базы. Но не смотря на то, что системы еще только набирают темп в своем развитие, они имеют большие перспективы.

Список литературы

[1] Михайлов В.Г., Златоустова Л.В. Измерение параметров речи. М.: Радио и связь, 1987. 168 с.

[2] Хлопенкова А.Ю., Белов Ю.С. Исследование алгоритмов автоматического распознавания речи на основе акустического и языкового моделирования // Научное обозрение. Технические науки, 2018, № 1. С. 32-36.

[3] Хорниг Н. Распознавание речи – задача не из легких. URL: https://newochem.io/voice-recognition/ (Дата обращения 30.03.2020)