Цифровая обработка сигналов

Лабораторная работа № 2

Основной тон голоса

Содержание

Организация:	Самарский университет
Подразделение:	Кафедра геоинформатики и информационной безопасности
Автор:	Юзькив Руслан Романович [yuzkiv.rr@ssau.ru]
Версия:	2022.09.28 (электронная)

Волк: Эй вы, серые козлята!
Ваша мать пришла, ребята!
Поскорей откройте дверь.
Я голодная, как зверь!
Молока полны копытца,
Стоит вам поторопиться!

Козлята: Не откроем волку дверь!
Уходи мохнатый зверь!
Песню ты не так поёшь,
Нас козлят не проведёшь!

Волк (зрителям): Что ж, пойду я к кузнецу –
Горло там перекую.
Песню тоненько спою,
Все козлят перехитрю!

Людмила Трифонова,
«Волк и семеро козлят»
по русской народной сказке

1 Теоретические сведения

1.1 Основной тон голоса

Частота основного тона — частота колебания голосовых связок при произнесении тоновых звуков. При выдохе воздух из лёгких проходит через голосовые связки, которые при этом начинают колебаться. Речевой тракт, состоящий из глотки, гортанной, оральной и носовой полостей, можно представить как нелинейный фильтр, который превращает основной тон в звуки речи (см. рисунок 1). Параметры такого фильтра непрерывно изменяются со временем и определяются согласованной работой большого числа мышц речевого тракта.

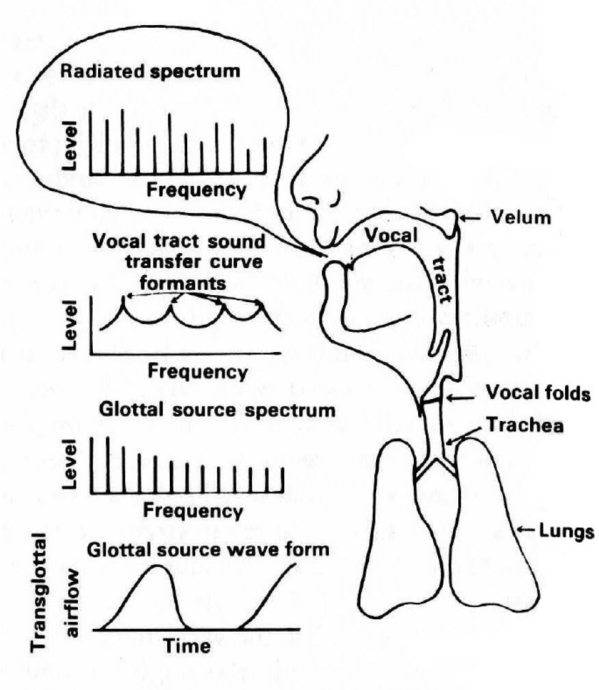

Рисунок 1 — Схема образования речевого звука

Источник: Oriol Nieto, Voice Transformations for Extreme Vocal Effects (рисунок 1.1)

Приведём некоторые из задач, для решения которых может использоваться значение частоты основного тона.

1. Определение пола человека. Частота основного тона мужского голоса лежит в диапазоне 80…240 Гц, женского — 140…500 Гц. Отметим, что это очень условные границы и значение может выходить за указанные пределы.

2. Распознавание эмоций. Частота основного тона непостоянна даже для отдельно взятого человека и зависит от его эмоционального состояния.

3. Сегментация аудио с несколькими голосами. Различие частот основного тона между разными людьми можно использовать для определения того, на каких участках аудиозаписи говорит один человек, а на каких — другой.

1.2 Автокорреляционная функция (АКФ)

В статистике автокорреляционная функция (АКФ) случайного процесса описывает корреляцию (сходство) между значениями процесса в различные моменты времени:

$$R_x(t,s) = E\left[(x(t) - \mu_x(t)) \cdot \overline{(x(s) - \mu_x(s))}\right],$$

где t, s — моменты времени (вещественные для непрерывного процесса и целые для дискретного), E – оператор математического ожидания, $x(t)$ — значение случайного процесса в момент времени t, $\mu_x(t)$ — среднее значение случайного процесса в момент времени t.

Для стационарных случайных процессов статистические характеристики не изменяются со временем. В частности, $\mu_x(t) = \mu_x$ (функция вырождается в константу), а значение АКФ не зависит от абсолютных значений моментов времени t и s, а зависит только от их разности $t - s$.

Кроме того, если процесс является эргодическим, то оператор усреднения E по всем реализациям случайного процесса (по ансамблю) может быть заменён усреднением по времени в пределах одной, достаточно продолжительной, реализации.

Тогда для дискретного вещественного сигнала $x[n]$ длины N, рассматриваемого как реализацию стационарного эргодического случайного процесса, автокорреляция может быть оценена по формуле

$$\hat{R}_x[m] = \frac{1}{N-m} \sum_{k=0}^{N-m-1} (x[k] - \mu_x) \cdot (x[k+m] - \mu_x)\,.$$

Среднее значение сигнала μ_x может быть оценено выборочным средним:

$$\hat{\mu}_x = \frac{1}{N} \sum_{k=0}^{N-1} x[k].$$

Отметим, что речь по своей природе не является стационарным процессом. Однако если сигнал рассматривать небольшими блоками длительностью порядка 20 мс, то на таких участках сигнал можно считать примерно стационарным. Учитывая, что основной тон речи в целом слабо меняется с течением времени, то в принципе для задачи оценки его частоты можно взять и намного больший участок сигнала.

Если частоту основного тона обозначить как f [Гц], то на небольшом участке сигнала должна явно просматриваться доминирующая периодичность с периодом $T = \frac{1}{f}$ [с]. Тогда при смещении такого сигнала на величину T он «более-менее совпадёт с самим собой» (грубо!) и даст пик на АКФ. Также пики должны наблюдаться

при смещении на $2T$, $3T$ и т. д. Таким образом, координата первого пика m_{max} функции $\hat{R}_x(m)$ соответствует периоду основного тона, выраженному в количестве отсчётов. Для нахождения частоты основного тона в герцах остаётся лишь перевести это значение в секунды (понадобится знание шага дискретизации) и вычислить обратную величину.

1.3 Преобразование Фурье в дискретном времени (ДВПФ)

Прямое преобразование Фурье в дискретном времени (ДВПФ):

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n]\ e^{-j\omega n}.$$

Так как комплексная экспонента $e^{-j\omega n}$ является периодической по аргументу ω с периодом 2π, то любая сумма таких экспонент тоже будет периодической. Значит, функция $X(e^{j\omega})$ — периодическая с периодом 2π.

Обратное преобразование Фурье в дискретном времени (ОДВПФ):

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})\ e^{j\omega n}\ d\omega.$$

Формула ОДВПФ лучше всего показывает суть преобразования — разложение дискретного сигнала $x[n]$ на линейную комбинацию комплексных экспонент $e^{j\omega n}$. Так как комплексная экспонента рассматривается как непрерывная по аргументу ω, то таких экспонент — несчётное множество, поэтому в записи линейной комбинации вместо суммы возникает знак интеграла. Комплексные коэффициенты разложения $X(e^{j\omega})$ называют непрерывным спектром сигнала $x[n]$.

Возвращаясь к задаче оценке основного тона голоса. Если ω — относительная циклическая частота основного тона, то в сигнале должны доминировать гармоники ω, 2ω, 3ω и т. д. С учётом представления гармонического сигнала в виде комплексных экспонент $\cos(\omega n) = \dfrac{e^{j\omega n} + e^{-j\omega n}}{2}$ получаем, что в амплитудном спектре $|X(e^{j\omega})|$ следует ожидать пики на частотах $\pm\omega$, $\pm 2\omega$, $\pm 3\omega$ и т. д. Таким образом, частота основного тона может быть оценена по позиции первого пика в непрерывном амплитудном спектре сигнала.

1.4 Алгоритм PSOLA

Алгоритм PSOLA предназначен для изменения частоты основного тона. Алгоритм состоит из двух основных этапов (рисунок 2). Обратите внимание, что алгоритм изменяет общую длительность сигнала.

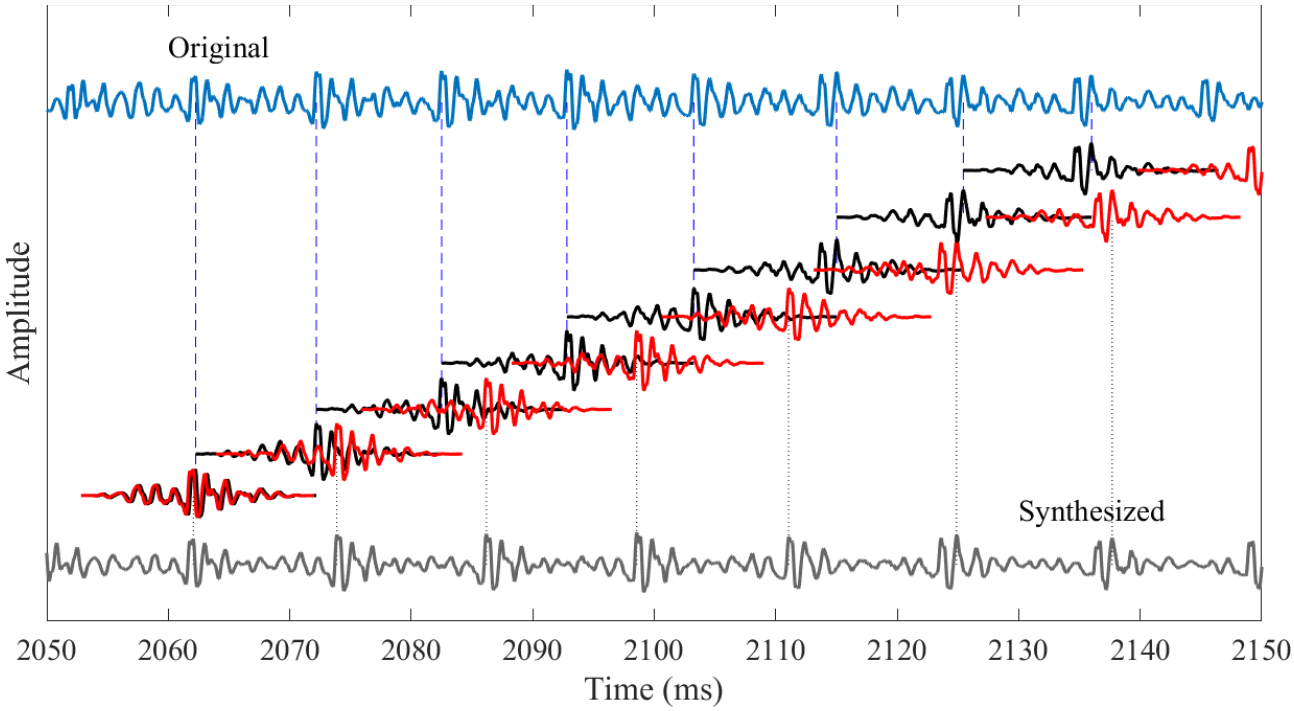

Рисунок 2 — Алгоритм PSOLA

Источник: Aalto University, Introduction to Speech Processing (раздел 3.13)

Этап 1. Сигнал «нарезается» на перекрывающиеся внахлёст кусочки длительностью $2T$ с шагом T, где T – период основного тона речи. По возможности каждый кусочек должен выбираться так, чтобы его центр приходился на пик амплитуды основной частоты. Это означает, что шаг T не является постоянным и может плавать в незначительных пределах. В рамках данной работы для определения центра каждого такого кусочка предлагается использовать библиотеку Google REAPER (Robust Epoch And Pitch Estimator), пример использования которой приведён далее в практическом подразделе. Однако попытки написать и использовать собственный алгоритм не воспрещаются и приветствуются 😂

Этап 2. Нарезанные кусочки суммируются обратно с изменённым расстоянием kT, где k — коэффициент изменения основной частоты голоса. Данный коэффициент является параметром алгоритма и задаёт во сколько раз нужно изменить основную частоту речи. Для подавления артефактов, которые неминуемо возникнут из-за перекрытия внахлёст, каждый кусочек нужно почленно умножить на оконную функцию, которая линейно «прибивает» сигнал к нулю к границам кусочка.

2 Практические сведения

2.1 Пакеты и функции

Пакет	Функция
`numpy`	`numpy.arange`
	`numpy.astype`
	`numpy.dot`
	`numpy.exp`
	`numpy.mean`
	`numpy.multiply`
`pyreaper`	`pyreaper.reaper`
`scipy`	`scipy.io.wavfile.read`
	`scipy.signal.windows.triang`
`statsmodels`	`statsmodels.tsa.stattools.acf`

2.2 Пример использования Google REAPER

```
import matplotlib.pyplot as plt
import numpy as np
import pyreaper
import scipy.io.wavfile as wavfile


# Загрузка и нормировка входного сигнала
fs, x = wavfile.read('input.wav')
x = x.astype(np.float32) / max(abs(min(x)), abs(max(x)))
t = np.linspace(0, (len(x) - 1) / fs, len(x))

# Подготовка данных для reaper
int16_info = np.iinfo(np.int16)
x = x * min(int16_info.min, int16_info.max)
x = x.astype(np.int16)

# Вызов reaper
pm_times, pm, f_times, f, _ = pyreaper.reaper(x, fs)

# Отображение позиций пиков
plt.figure('[Reaper] Pitch Marks')
plt.plot(t, x)
plt.scatter(pm_times[pm == 1], x[(pm_times * fs).astype(int)][pm == 1], marker='x', color='red')

# Отображение значений основной частоты
plt.figure('[Reaper] Fundamental Frequency')
plt.plot(f_times, f)

print('Average fundamental frequency:', np.mean(f[f != -1]))

plt.show()
```

3 Задание на лабораторную работу

3.1 Подготовка входных данных

С помощью микрофона запишите короткий отрывок речи (одно-два предложения). Можно взять часть стиха, поэмы или какую-либо цитату.

На выходе должен получиться одноканальный `wav`-файл. Если ваша программа аудиозаписи не поддерживает `wav`, получившийся файл можно преобразовать с помощью какого-либо конвертера (например, VLC).

Если не удаётся получить одноканальный файл, то проблему можно решить на уровне кода скрипта на Python'е — при загрузке данных будет возвращен набор каналов (например, два для стерео), из которых для дальнейшей обработки возьмите любой один из каналов.

Все дальнейшие задания выполняются над полученным `wav`-файлом.

3.2 Оценка основного тона на основе АКФ

Шаг 1. Реализуйте вспомогательную функцию `my_acf` для вычисления значения АКФ $\hat{R}_x[m]$ при заданном аргументе. Функция должна принимать сигнал `x` и значение аргумента `m`; возвращать — посчитанное значение АКФ для заданного аргумента.

Шаг 2. Проверьте правильность реализации, сравнив результаты работы своей функции для различных `m` с результатами работы библиотечной функции `statsmodels.tsa.stattools.acf` (эту функцию необходимо использовать с аргументом `adjusted=True`). Обратите внимание, что библиотечная функция рассчитывает АКФ сразу для множества значений `m`, т. е. библиотечную функцию достаточно вызвать один раз. Так как реализация `my_acf` «в лоб» может работать очень медленно, используйте для экспериментов короткий сегмент сигнала.

После проверки своей реализации на корректность для дальнейших вычислений используйте только библиотечную функцию как более быструю в вычислительном плане.

Шаг 3. Постройте график АКФ и оцените по нему частоту основного тона речи.

3.3 Оценка основного тона на основе ДВПФ

Шаг 1. Реализуйте вспомогательную функцию `my_dtft` для вычисления значения дискретного во времени преобразования Фурье. Функция должна принимать сигнал `x`, частоту дискретизации `fs` (в герцах; понадобится для пересчёта между различными видами частот) и значение частоты `f` (в герцах), для которой нужно вычис-

лить значение спектра; возвращать — посчитанное значение амплитудного спектра. Если в качестве `f` передан итерабельный объект (т. е. не скаляр, а вектор частот), то функция должна вычислить значение спектра для каждого значения входной частоты и вернуть `numpy`-массив соответствующих значений амплитудного спектра.

Для ускорения расчётов необходимо реализовать формулу ДВПФ в виде скалярного произведения двух векторов, а не в виде цикла. То есть необходимо подготовить вектор, соответствующий $e^{-j\omega n}$ (`numpy.arange`, `numpy.exp`), а затем найти скалярное произведение данного вектора с анализируемым сигналом (`numpy.dot`). Получившееся комплексное значение нужно взять по модулю — это и будет амплитудное значение спектра.

Шаг 2. Используя функцию `my_dtft`, постройте график амплитудного спектра в адекватной полосе частот (например, от 40 до 500 Гц с шагом 1 Гц) и оцените по нему частоту основного тона речи.

3.4 Оценка основного тона с помощью Google REAPER

Оцените основной тон речи с помощью библиотеки Google REAPER. В качестве основы см. пример.

3.5 Изменение основного тона на основе алгоритма PSOLA

Шаг 1. Реализуйте функцию `psola` для изменения тона речи согласно алгоритму, описанному в разделе Алгоритм PSOLA. Функция должна принимать сигнал `x`, частоту дискретизации `fs` (в герцах) и коэффициент `k` изменения основной частоты голоса. Функция должна вернуть сигнал с изменённой частотой основного тона речи.

Алгоритм рекомендуется (но не требуется) применять только к тем участкам сигнала, где есть тоновые звуки. Используйте библиотеку Google REAPER для определения тоновых участков речи, центров нарезаемых сегментов и длины сегмента. Напомним, что длина сегмента определяется из оценённой частоты основного тона.

Шаг 2. Измените частоту основного тона речи с помощью написанной функции `psola` и запишите результат в `wav`-файл.

Выполнение последнего шага допускается в команде в виде творческого диалога (например, разыграть какой-нибудь анекдот), когда голос каждого участника (или даже одного участника в разные моменты времени) изменяется со своим набором параметров. Например, для одного участника диалога частота основного тона повышается, для другого — уменьшается.

4 Формат сдачи

Предоставить скрипт, входной `wav`-файл, таблицу с оценками основной частоты голоса (в произвольном формате в виде простого `txt`-файла) и выходной `wav`-файл с изменённым основным тоном.

Скрипт должен содержать:

1) реализованную функцию `my_acf`;

2) тест, демонстрирующий эквивалентность результатов работы функций `my_acf` и `statsmodels.tsa.stattools.acf`;

3) реализованную функцию `my_dtft`;

4) реализованную функцию `psola`;

5) код, который строит графики АКФ и ДВПФ (по которым проводилась оценка основного тона) и визуализирует результаты работы Google REAPER;

6) код генерации выходного `wav`-файла.

Таблица с оценками основной частоты голоса должна содержать:

1) оценку, полученную с помощью АКФ;

2) оценку, полученную с помощью ДВПФ;

3) оценку, полученную с помощью Google REAPER.

5 Контрольные вопросы

1. Определение АКФ и ДВПФ.

2. Чему равно значение ДВПФ для сигнала $\{1, 2, 2, 1\}$ при $\omega = 0$ и при $\omega = \pi$? Вычислите эти значения вручную на бумаге, не используя каких-либо вычислительных средств.

3. С учётом периодичности ДВПФ есть смысл вычислять значения только для какого-то одного периода, например при $\omega \in (-\pi, \pi]$ $\text{рад}/\text{отсчёт}$. Если сигнал является вещественным, то спектр будет обладать чётной симметрией — информативным останется только полупериод $\omega \in [0, \pi]$ $\text{рад}/\text{отсчёт}$. Для какой максимальной частоты (в герцах) есть смысл вычислять ДВПФ, если $f_s = 44100$ Гц? Другими словами, начиная с какой частоты значения спектра перестанут быть информативными и начнут повторяться?

4. С помощью функции из прошлой лабораторной создайте чистый тон с некоторой частотой f и длительностью 1 с. Постройте график спектра такого сигнала с помощью `my_dtft` с достаточно мелким шагом в области частоты f. Почему на построенном графике виден не строгий чёткий пик при соответствующий частоте тона, а некоторая его размытая версия?