СИНТЕЗ ФРАГМЕНТОВ ГОЛОСА ЧЕЛОВЕКА НА ОСНОВЕ РЕКОНСТРУКЦИИ ЧАСТОТНЫХ СПЕКТРОВ

Г. Лань, А.С. Фадеев, А.Н. Моргунов
Научный руководитель: А.С. Фадеев
Томский политехнический университет
E-mail: lanber@tpu.ru

Введение

Одним из основных аспектов синтезирования человеческой речи является – качество выходного сигнала, от чего зависит пригодность использования данной технологии в коммерческих целях. TTS-алгоритмы (Text to Speech) являются самым распространенным применением данной технологии [1, 2]. Основная задача в разработке систем синтеза речи – разработать систему и методику генерации человеческой речи, результаты которой будут таковы, что человек на слух не смог бы отличить сгенерированный сигнал от того, что был записан реальным индивидом.

В настоящей работе описана методика синтеза отдельных временных отрезков сигнал на основе сведений о частотном спектре звукозаписей речи человека.

Синтез фонем звуков

Согласно [4] гласный звук состоит из одного слога, слог состоит из одной или нескольких фонем. Фонемы являются наименьшими элементарными составляющими речи человека, синтез которых на основе аналитических моделей и наборов связных параметров позволит генерировать более крупные элементы речи: звуки, буквы и слова.

Для упрощения моделирования и синтеза отдельных формант можно применить следующую аналитическую модель на квазистационарных участках:

$$\varphi_i(t) = A_i \sin(2\pi\nu_i t) \quad (1)$$

где A_i и ν_i – амплитуда форманты и частота форманты i соответственно. Аналитическая модель квазистационарного участка фонемы будет иметь вид:

$$f(t) = \sum_i A_i \sin(2\pi\nu_i t). \quad (2)$$

Для определения параметров A_i и ν_i, а также числа отдельных формант, используется оконное преобразование Фурье (ОПФ) и полученная на его основе спектрограмма записанного сигнала.

Спектрограмма звукового сигнала

Основным шагом разработанной методики является получение спектрограммы анализируемого сигнала, которая является результатом применения ОПФ:

$$STFT(\tau, \omega) = \int_{-\infty}^{+\infty} f(t) W(\tau - t) e^{-j2\pi\nu\tau} d\tau \quad (3)$$

Здесь $W(\tau - t)$ – оконная функция, $f(t)$ – значение амплитуды исходного сигнала в момент времени t, τ – положение окна, f – частота компоненты. В виду малой продолжительности по времени записанного сигнала, окно ОПФ было подобрано таким образом, чтобы в результате получить высокое частотное разрешение.

Квазистационарное состояние

Известно, что процесс воспроизведения звучания человеческого голоса нестабильный и непериодический [8], в данной работе анализируется только квазистационарный временной отрезок звучания фонемы (период выдержки). Значения параметров фонемы и их точность зависят от свойств выбранного временного отрезка. Анализ спектрограмм фонем гласных букв (рисунок 1) показывает, что несмотря на не стационарность всего сигнала и каждой форманты по отдельности, отдельные короткие временные отрезки, продолжительностью $10^{-1} - 10^{-2}$с. являются квазистационарными и для них возможно применение отписанной модели.

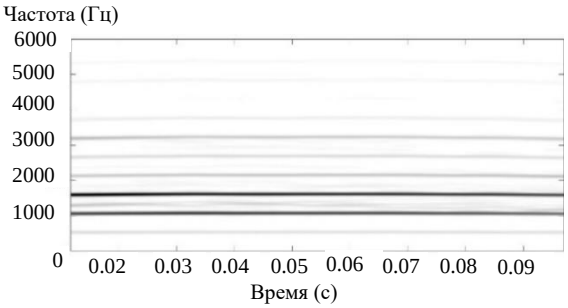

Рис. 1. Спектрограмма квазистационарного участка «А»

Получение спектра

Для построения спектра сигнала применено быстрое преобразование Фурье (БПФ) в пакете прикладных программ MATLAB:

$$F_k = \sum_{j=1}^{N} f_j W_N^{(j-1)(k-1)} \quad (4)$$

где $W_N = e^{(-2\pi i)/N}$ – комплексная экспонента, f_j – исходный сигнал, j, k – номер отсчета в последовательностях f_j и F_k соответственно, N – количество отсчетов.

Вычисление коэффициентов для отдельных формант

На следующем этапе были получены наборы векторов F_i, содержащих параметры P_{ij}, которые необходимы для синтеза отдельной форманты. Такие вектора составляют матрицу параметров M_Φ формант записанного звукового сигнала.

$$M_\Phi = [F_1 \quad F_2 \quad F_3 \quad \cdots \quad F_i \quad \cdots \quad F_n] \quad (5)$$

$$F_i = [P_{i1} \quad P_{i2} \quad P_{i3} \quad \cdots \quad P_{ij} \quad \cdots \quad P_{im}] \quad (6)$$

Для построения математической модели и реконструкции отдельных формант был предложен метод параметрического описания формант речи

человека на основе полученных частотно-временных характеристик [3]. Очевидно, что амплитуда A_i и частота V_i являются основными параметрами, характеризующими поведение каждой форманты гласного звука человеческой речи во времени.

Генерация сигнала и оценка качества синтезированного сигнала

В настоящей работе оценка качества предложенной методики для генерации звукового сигнала, имитирующего голос человека, производится на основе полученных параметров матриц.

$f(t) = \sum_{i=1}^{n} A_i \cdot \sin(2\pi\nu_i \cdot t)$, (6)

где $\sin(2\pi\nu_i \cdot t)$ – форманта сигнала, ν_i – частота форманты, A_i – значение амплитуды.

Качество синтетического сигнала оценивается путем сравнения АЧХ полученного сигнала и исходного. Для большей наглядности приведенные спектры показаны в логарифмических осях на рисунке 2. Пунктирная линия представляет синтезированный сигнал.

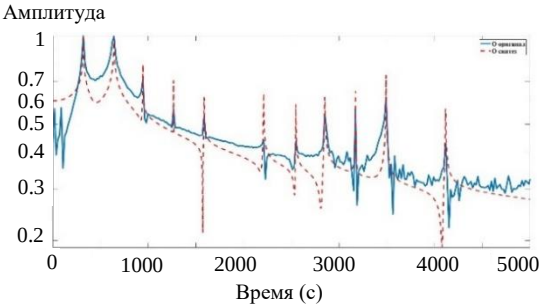

Рис. 2. Спектры звукового сигнала «О» (логарифм)

Анализ изображения показывает, что спектры оригинального и синтезированного на основе полученных коэффициентов сигналов практически совпадают.

Особенности синтеза. Отличия гласных звуков

В настоящем исследовании были рассмотрены различные гласные звуки. На рисунке 3 показаны спектрограммы синтезированных звуков из разного набора формант для одной и той же гласной буквы произнесенной диктором в разных тональностях. Анализ изображений спектрограмм показывает, что одна и та же гласная буква имеет схожие отношения частот фонем, но их амплитуда и абсолютное значение частоты сильно зависят от относительного тона произношения этой буквы диктором.

На рисунке 4 показано отношение значения частот формант гласных звуков ***А*** и ***О*** к частоте основной форманты с разным набором формант: без эмоциональной окраски (data1), относительно низкая частота (data2), относительно высокая частота (data4) и естественные частоты голоса (data3). Качественно представленные тренды имеют одинаковую форму и поведение.

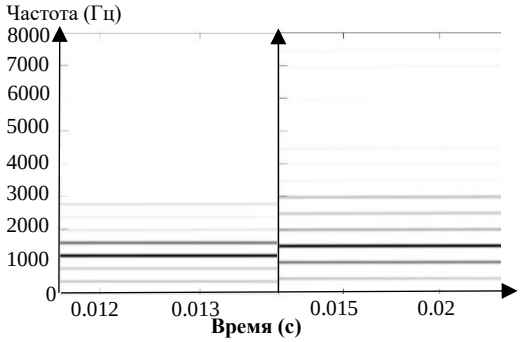

Рис. 3. Спектрограммы буквы «А» на относительно низких частотах (левый) и на естественных частотах (правый)

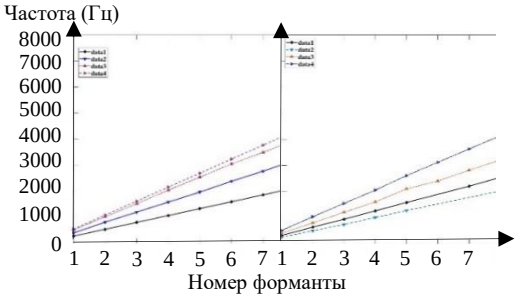

Рис. 4. Кривая изменений частот звука «А» (левый) и «О» (правый)

Заключение

В данной работе предложена методика распознавания отдельных фонем некоторых букв человека, при помощи спектра и спектрограммы оригинальных фонем получены основные амплитудно-частотные характеристики речи человека. Предложена методика, позволяющая синтезировать сигнал фонем человека на основе полученных параметров, а также методика оценки качества синтезированных элементов речи. В работе также приведены некоторые выявленные особенности частотных характеристик наборов формант для букв, произнесенных на разных относительных частотах.

Список использованных источников

1. Синтез речи: Википедия – свободная энциклопедия [электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Синтез_речи.
2. Рыбин С. В. Синтез речи. Учебное пособие. — СПб.: Университет ИТМО (2014). — С. 21 – 25.
3. Способы оценки субъективного качества речи [электронный ресурс]. — URL: https://habr.com/en/post/177099/.
4. Phoneme (Lexicon of Linguistics). Universiteit Utrecht. [2014-11-03].
5. Лань Г., Моргунов А.Н., Методика реконструкции фонем голоса человека // Вестник современных исследований. — Выпуск № 10-3 (25) 2018, ISSN 2541-8300. — С. 130 – 135.