Принцип формирования тестовых наборов сигналов в речевом диапазоне частот для исследования алгоритмов обработки аудиозаписей

У. В. Токарева[1], А. Д. Шульженко[2]

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[1]tokareva.ulv@gmail.com, [2]adshulzhenko@etu.ru

Аннотация. **В статье рассматривается разработка метода для формирования тестовых наборов сигналов в речевом диапазоне частот. В работе представлены рассмотрение существующих математических моделей генерации сигналов, а также вывод модернизированной формулы, предназначенной для формирования сигнала, более приближенного к реальной голосовой записи. На основании формул, полученных в результате данной работы, был реализован программный код для построения тестовых сигналов и сравнения искусственного и реального сигналов между собой.**

Ключевые слова: форманты; модуляция; среднеквадратическое отклонение; тестовые наборы; синтез речевых сигналов; цифровая обработка речи

I. Введение

Цифровая обработка речи играет ключевую роль в различных областях, включая телекоммуникации, биометрию и распознавание речи. Однако для создания эффективных методик обработки голосовых данных и для исследования алгоритмов обработки аудиозаписей нужно иметь исчерпывающий набор тестовых сигналов в речевом диапазоне частот.

В настоящее время большое количество работ посвящено анализу тестовых наборов аудиозаписей. Традиционно для формирования тестовых наборов сигналов в речевом диапазоне частот использовались реальные голосовые записи, однако данный способ формирования выборки имеет ряд проблем, так как «записи речи часто содержат фоновый шум, реверберацию, снижение частотной полосы и другие искажения» [1]. Также среди ограничений использования набора записей реальных голосов можно отметить сложность в создании разнообразных условий записи и необходимость большого объема данных. Важно учитывать различные условия записи и их влияние на качество речевых сигналов [2]. Таким образом, мы можем сделать вывод, что использование набора реальных голосовых записей не подходит для полноценного анализа и исследования алгоритмов обработки аудиозаписей.

В качестве подхода к решению данной проблемы можно выделить генерацию собственных наборов аудиофайлов, которые будут обладать схожими характеристиками с голосовыми аудиозаписями. Если говорить об актуальности генерации собственных тестовых наборов, то в современных работах рассматривается реконструированная модель речевого процесса, а также возможность использования различных методов численного интегрирования для восстановления речевого сигнала [3] Также в современных работах рассмотрены теоретические основы цифровой обработки речевых сигналов, в том числе их физические свойства и цифровое представление, спектральный и корреляционный анализ [4]. В статье [5] указано, что существуют методы синтеза речевых сигналов, которые базируются на моделях, аппроксимирующих голосовые сигналы в виде полигармонических колебаний с амплитудной и частотной модуляцией. Также в указанной выше статье рассмотрен инструментарий для проведения исследований речевых сигналов и создания алгоритмов обработки распознавания речи. Это позволяет создать тестовые наборы, обладающие спектральными характеристиками, аналогичными реальным голосовым записям.

Таким образом, мы видим, что данная тема имеет широкое распространение среди современных исследований, но также мы можем сделать вывод, что готовые наборы аудиозаписей не подходят для исследования, так как они обладают рядом ограничений, среди которых выделяются следующие: ограниченное разнообразие условий записи, проблемы с конфиденциальностью авторов записей, невозможность точного контроля характеристик сигнала, отсутствие универсальности и ограниченность объемов данных.

В настоящей работе решается задача разработки и обоснования метода формирования тестовых наборов сигналов в речевом диапазоне частит для исследования алгоритмов обработки аудиозаписей.

II. Существующие решения

Голосовой сигнал является сложным акустическим явлением и состоит из двух видов сегментов – вокализированных и невокализированных. Вокализированные сегменты могут быть описаны как полигармонические сигналы, в то время как шумовые составляющие (невокализированные сегменты) можно описать как белый шум.

Для описания вокализованных сегментов речи используется математическую модель вокализованного сегмента речи y(t), являющуюся решением ДУ, которое описывает прохождение периодического колебания от источника в виде голосовых связок (т.е. полигармонического сигнала или нескольких гармоник

ряда Фурье) через систему параллельных резонаторов с затуханием [6]:

$$u(t, f_0) = \sum_{l=1}^{L} U_l \cos\left[2\pi l f_0 t + l \cdot m \cdot \sin(2\pi F_0 t)\, \varphi_l\right], \quad (1)$$

где U_l – амплитуда l-й гармоники к несущему колебанию; f_0 – частота основного тона; F_0 – наименьшая частота модулирующего колебания; φ_l – начальная фаза l-й гармоники к несущему колебанию; L – количество несущих гармоник; t – длительность вокализированного сегмента; m – частотная модуляция.

Невокализированные сегменты речи не имеют периодической структуры, но содержат турбулентный шум, поэтому описываются с помощью другого подхода. Для более реалистичного моделирования речевых звуков часто используется фильтрованный белый шум:

$$n(t) = \int_{-\infty}^{\infty} h(\tau) \cdot w(t - \tau)\, d\tau, \quad (2)$$

где n(t) – сигнал невокализированного сегмента речи; h(t) – импульсная характеристика фильтра, моделирующего резонансы вокального тракта; w(t) – белый шум, представляющий турбулентный шумовой источник.

Таким образом, общую формулу речевого сигнала можно представить в виде следующей математической модели:

$$S(t, f_0) = u(t, f_0) + n(t), \quad (3)$$

Для генерации сигнала необходимо учесть следующие параметры [7]:

- частота основного тона;
- шум;
- модуляция;
- количество формант;
- центральная чистота каждой форманты.

Таким образом, необходимо составить математическое описание данных аспектов для реализации генерации сигнала в речевом диапазоне частот.

Для синтеза звуков могут применяться три основных типа модуляции: амплитудная, частотная и периодическое изменение гармоник, содержащихся в звуке. Применение модуляции для синтеза позволяет получить в результирующем выходном сигнале значительно большее количество спектральных составляющих, которые влияют на естественность звучания. Образование шума представляет собой достаточно сложный процесс, зависящий от многих факторов, поэтому для моделирования шума в качестве альтернативы обычно используют белый шум, спектр которого распределен по некоторому закону относительно некоторой центральной частоты.

Таким образом, мы видим, что представленные выше формулы не обладают достаточной полнотой для генерации сигнала в речевом диапазоне частот, так как в них не учитывается амплитудная модуляция, нет описания формант. По этой причине для моделирования вокализированной части сигнала в речевом диапазоне частот лучше использовать формулу, описанную в работе [8]:

$$u(t) = \sum_{k=0}^{K} M_k \cos(2\pi k F_0 t + \Phi_k) \sum_{l=1}^{L} U_l \cos[2\pi l f_0 t + l \cdot m_i \sin(2\pi F_0 t + \Psi_0) + \varphi_0], \quad (4)$$

где $t\epsilon[0; t_u], t_u$ – длительность звука; U_l – амплитуда гармоники колебания; $\sum_{k=0}^{K} M_k \cos(2\pi k F_0 t + \Phi_k)$ – составляющая, которая характеризует переходные процессы (нарастание и спад амплитуд сигнала) и экспоненциальное затухание составляющих сигнала; m_i – индекс модуляции.

III. Предлагаемое решение

Для программной реализации алгоритма формирования тестовых сигналов в речевом диапазоне частот необходимо несколько модернизировать формулу, описывающую вокализированные сегменты сигнала. Необходимо, чтобы модель учитывала нелинейные эффекты вокального тракта и динамику сигнала через затухание и модуляцию амплитуд. Изменения в реальной речи происходят плавно, поэтому необходимо реализовать интегрирование фаз и динамическую нормализацию амплитуд. С учетом перечисленных требований получаем новую математическую модель вокализированного сегмента сигнала:

$$u(t) = \sum_{k=0}^{K} M_k \cos(2\pi k F_0 t + \Phi_k) \sum_{l=1}^{L} U_l(t) \cos[2\pi l f_0 t + l \cdot m_i \int_0^t \sin(2\pi F_0 \tau + \Psi_0) d\tau + \varphi_0], \quad (5)$$

где $U_l(t) = U_l \cdot e^{-at} \cdot (1 + \beta \cos(2\pi F_0 t))$ – динамическая формула амплитуды с учетом экспоненциального затухания.

Данная формула учитывает форманты и амплитудно-частотную модуляцию, поэтому сигнал в речевом диапазоне частот будет формироваться по формуле (3), где шумовой компонент будет описан формулой (2), а вокализированный – формулой (5).

Графики спектров сгенерированных записей и голосов в этом же частотном диапазоне представлены на рис. 1.

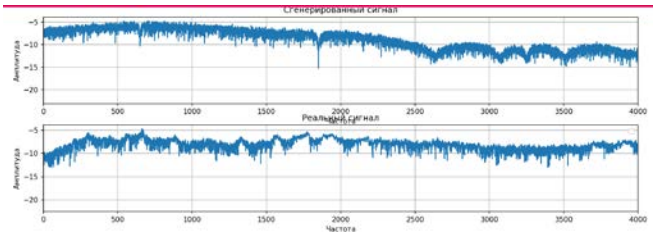

Рис. 1. Графики спектров сгенерированных записей и голосов в этом же частотном диапазон

IV. Реализация и тестирование

В ходе данной работы была разработана программа, предназначенная для генерации сигналов в речевом диапазоне частот с возможностью тонкой настройки параметров. Программный комплекс предусматривает возможность задавать такие параметры как: частота дискретизации, центральные частоты формант, параметры амплитудной модуляции (базовая амплитуда, глубина модуляции), параметры частотной модуляции (настройка вариации частоты формант п настройка плавного изменения тона), дополнительно задаются данные для уровня шума. Для доказательства правомерности замещения формул необходимо построить разницу спектров и линию тренда, а также рассчитать среднеквадратичное отклонение для каждой пары сигналов, где один сигнал является реальной голосовой записью, а второй – результатом формирования в рамках программы. Разные пары

сигналов иллюстрируют различные условия записи голоса.

Рассмотрим запись относительно громкого женского голоса, сделанную в помещении. Результат генерации для громкого женского голоса в помещении представлен на рис. 2.

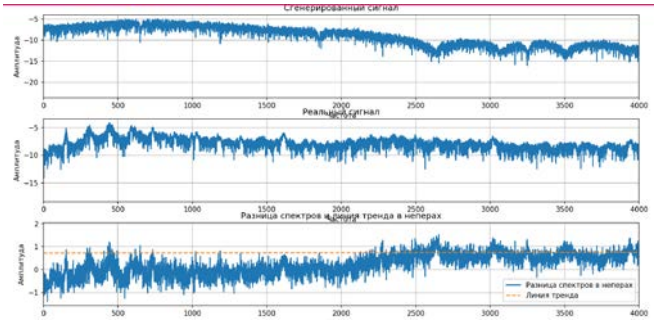

Рис. 2. Результат генерации для женского голоса в помещении

Также для данного случая среднеквадратичное отклонение равно 0.93013 Np.

Теперь попробуем смоделировать прерывистый шепот женским голосом. Результат генерации прерывистого шепота в тишине представлен на рис. 3.

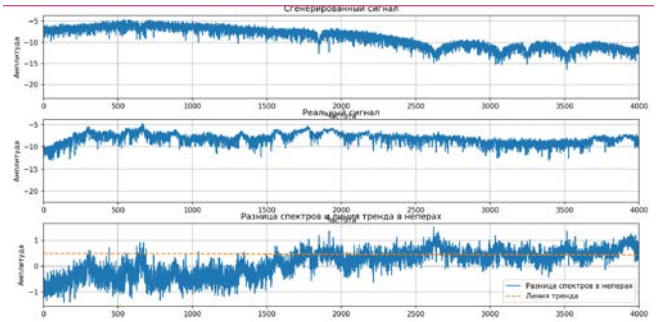

Рис. 3. Результат генерации прерывистого шепота в тишине

Среднеквадратичное отклонение равно 0.49562 Np.

Далее рассмотрим голосовое сообщение, записанное мужчиной в тихом помещении. Результат генерации мужского голоса представлен на рис. 4.

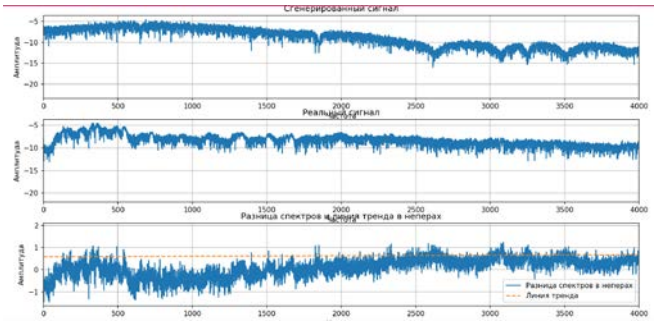

Рис. 4. Результат генерации мужского голоса

Среднеквадратичное отклонение равно 0.98764 Np.

Исходя из вышеописанных примеров видно, что настройка параметров генерации позволяет создать тестовые наборы, соответствующие по характеристикам реальным голосам, записанным в разных условиях. Также мы можем сделать вывод, что настройка получается достаточно точной, так как среднеквадратичное отклонение низкое.

V. Заключение

В настоящее время большое количество работ требует базу данных аудиосигналов, необходимую для проведения исследований. Так как использование реальных голосовых записей для формирования тестовых наборов данных имеет ряд ограничений и проблем, возникает необходимость генерации сигналов в речевом диапазоне частот. Основная задача – это применение в формировании тестовых наборов искусственных записей, которые по своим характеристикам совпадают с реальными.

В данной работе были рассмотрены существующие математические модели для формирования искусственных речевых сигналов и результаты формирования сигналов на основании данных математических формул.

Было предложено модифицировать существующие математические модели для большего приближения к реальным голосовым сигналам и на основании этой математической модели реализовать программный комплекс, позволяющий опытным путем проверить качество и схожесть сгенерированных сигналов.

В ходе работы был реализован способ формирования сигналов в речевом диапазоне частот с возможностью настройки под различные тембры и условия записи. Данным способ формирования основывается на доработанной математической модели.

Метод показал свою эффективность, однако в дальнейшем необходимо проработать возможность формирования большего количество записей и предусмотреть условия для более точной настройки, чтобы охватить все тонкости, которые содержат в себе аудиозаписи.

В качестве дальнейшего направления исследования рассматривается проработка метода более детального формирования тестовых сигналов в речевом диапазоне частот, которые будут максимально приближены по своим характеристикам к реальным голосовым записям, а также разработка способа автоматической генерации широкой и разноплановой выборки сигналов.

Список литературы

[1] Андреев П.К. Генеративные модели для улучшения речи. Автореф. дис. … канд.комп.наук. Москва, 2024.

[2] Козлачков С.Б., Дворянкин С.В., Бонч-Бруевич А.М. Принципы формирования тестовых речевых сигналов при оценках эффективности технологий шумоочистки // Вопросы кибербезопасности №3(27) – 2018.

[3] Копытов В.В., Якушев Д.В. Методы кодирования речевых сигналов с помощью реконструированной модели речевого процесса // Известия ЮФУ. Технические науки. С.37-44.

[4] Ахмад Х.М., Жирков В.Ф. Введение в цифровую обработку речевых сигналов. Владимир – 2007.

[5] Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я. Современный инструментарий для разработки речевых технологий // ИТиВС, 2004, выпуск 2, 11–24.

[6] Бакаев А.В. Влияние форматных областей на разборчивость речи // Информационное противодействие угрозам терроризма. 2008. №. 11. С. 83-90.

[7] Фролов А.В., Фролов Г.В. Синтез и распознавание речи. 2003. https://www.frolov-lib.ru/books/hi/ch00.html

[8] Гущина А.А. Разработка и совершенствование математических моделей речевых сигналов для задач анализа и синтеза речи: Автореф.дис. …канд.техн.наук. Воронеж. 2014.