Вейвлет–анализ аудиосигналов и синтез речи

С. А. Никоноров* А. Н. Боголюбов
Московский государственный университет имени М. В. Ломоносова, физический факультет
Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2
(Статья поступила 25.06.2018; Подписана в печать 10.09.2018)

Разработан алгоритм анализа и аппроксимации речевых сигналов с использованием непрерывного вейвлет–преобразования. В данной работе были исследованы аудиофайлы голосовой базы Zuiga Mizuki, некоторые из них (файлы гласных звуков) были восстановлены при помощи предложенного алгоритма синтеза.

PACS: 02.60.Gf УДК: 519.6
Ключевые слова: вейвлет-преобразование, анализ речевых аудиосигналов, алгоритм синтеза речи, алгоритм трассировки скелетона.

ВВЕДЕНИЕ

Необходимость анализа и синтеза сигналов разного рода возникла достаточно давно и актуальна до сих пор [1]. Для этого применяется большое количество различных методов, среди которых нельзя не выделить Фурье–анализ как один из наиболее общих методов анализа и обработки сигналов. Однако для изучения процессов с меняющимися во времени характеристиками более подходящим инструментом является вейвлет–анализ [2–4], который лежит в основе идеи алгоритма анализа и синтеза аудиосигналов, представленного в данной работе.

Задача синтеза речи была поставлена достаточно давно, однако она актуальна и на сегодняшний день: установлено, что речь является для человека наиболее удобным и естественным способом обмена информацией. Это означает, что при таком способе человек допускает меньше ошибок, меньше устает, быстрее реагирует, а скорость обмена информацией выше, чем при других способах — визуальном, тактильном, тонально–звуковом [5]. Поэтому, когда возникает необходимость в обмене информацией между человеком и, например, ЭВМ, на настоящий момент используются не только визуальные и/или тактильные средства (монитор, клавиатура, мышь), но и голосовые (системы распознавания и синтеза речи) [6].

В данной работе представлен алгоритм синусоидально–конкатенативного синтеза речи, который можно разбить на:

1. спектрально-временной анализ речевого аудиосигнала;
2. оптимизацию результатов анализа:
3. отсечение несущественных гармоник;
4. аппроксимация временных зависимостей оставшихся компонент сигнала;
5. конструктивный синтез речевого аудиосигнала.

Синтез сигнала в данном случае выполняется посредством генерации синусоидальных волн, зависимость частоты и амплитуды от времени которых определяется в результате анализа исходного речевого сигнала.

В данной работе спектрально-временной анализ речевых аудиосигналов выполнялся посредством непрерывного вейвлет–преобразования. Данное преобразование также как и Фурье–преобразование позволяет выполнить разложение исследуемого сигнала по некоторому семейству анализирующих функций. Однако в данном случае:

1. класс допустимых анализирующих функций гораздо шире, чем в случае Фурье–преобразования;
2. при проведении вейвлет-преобразования у анализирующих функций меняется как «частота» (масштаб) так и «положение на временной оси» (сдвиг аргумента по оси времени), что позволяет проводить так называемый кратномасштабный анализ.

Изменение «частоты» позволяет исследовать особенности сигнала разного масштаба (с разной степенью «детализации»), а изменение сдвига аргумента по времени позволяет исследовать его на разных временных интервалах. Данное свойство вейвлетов позволяет им «сфокусироваться» на сингулярностях или резких перепадах сигнала, тогда как оконные преобразования Фурье для этого не подходят [7].

Рассмотрим процесс построения матрицы преобразования. Пусть $w(t)$ — выбранный анализирующий вейвлет. Он порождает семейство функций $w_{n,m}(t)$:

$$w\left(t\right) \rightarrow \ w_{n,m}\left(t\right) = w\left(nt+m\right).$$

Будем называть m параметром сдвига, n — параметром растяжения.

В дальнейшем будем рассматривать только сдвиги и растяжения следующего вида:

$$m = -\frac{\varepsilon_x x}{\sigma^{\varepsilon_y y}}, \quad n = \frac{1}{\sigma^{\varepsilon_y y}}$$

*E-mail: nics1543@gmail.com

где

$$x, y \in \mathbb{Z}; \quad \varepsilon_x, \varepsilon_y, \sigma \in \mathbb{R}.$$

Определим покомпонентно матрицу коэффициентов непрерывного вейвлет–преобразования C:

$$C_{n,m} = \int_{-\infty}^{+\infty} F(t) \cdot w_{n,m}(t) dt.$$

С учетом введенных ограничений на параметры n и m получим:

$$C_{x,y} = \int_{-\infty}^{+\infty} F(t) \cdot w\left(\frac{t - \varepsilon_x x}{\sigma^{\varepsilon_y y}}\right) dt.$$

C хранит информацию о временной зависимости амплитуды, частоты и фазы каждой компоненты сигнала $F(t)$.

Заметим, что матрица вейвлет–преобразования, как и спектр, полученный посредством оконного преобразования Фурье, также может содержать (и, как правило, содержит) так называемые «артефакты». Однако класс анализирующих вейвлетов достаточно широк, что позволяет выбирать оптимальные функции для каждой конкретной задачи.

В данной работе для анализа сигналов использовалось следующее семейство функций:

$$w_{a,b}^{x,y;\varepsilon_x,\varepsilon_y,\sigma}(t) = \sin\left(a\pi \frac{t - \varepsilon_x x}{\sigma^{\varepsilon_y y}}\right) e^{-\left(\frac{b(t-\varepsilon_x x)}{\sqrt{2}\sigma^{\varepsilon_y y}}\right)^2}.$$

Обозначения: a, b — параметры вейвлета; $\varepsilon_x, \varepsilon_y, \sigma$ — параметры дискретизации коэффициентов масштабирования/сдвига; x, y — коэффициенты масштабирования/сдвига.

Функции данного семейства представляют собой отмасштабированную и смещенную с различными коэффициентами мнимую часть вейвлета Морле. Данный вид вейвлета выбран по следующим причинам:

1. при анализе синусоидальных сигналов уровень артефактов, порожденных преобразованием с использованием данного семейства анализирующих функций, мал относительно «полезной» информации о сигнале;

2. для данного вейвлета легко рассчитать соответствие между частотой гармоники в герцах и координатой на соответствующей оси скейлограммы.

1. АЛГОРИТМ СИНТЕЗА РЕЧИ

Как было отмечено во введении, существует несколько подходов к решению задачи синтеза речи. Подход, основанный на «склеивании» заранее записанных слов и фраз дает высокое качество речи и в наши дни весьма распространен (объявления на вокзалах, аэропортах, метро и т.д.). Однако системы, реализующие данный метод синтеза, способны воспроизводить весьма ограниченный набор фраз. Для решения задачи «озвучивания» произвольного текста наиболее распространены метод конкатенации и метод полного синтеза речи по правилам, который в свою очередь может быть реализован несколькими способами. Метод конкатенации представляется достаточно очевидным и естественным — он заключается в «склеивании» полуслогов (полуслог — это сочетание целого согласного и половины гласного звука). Каждый полуслог записывается в виде аудиофайла и сохраняется в базу. При синтезе по введенному тексту каждому слогу сопоставляется соответствующий аудиофайл из базы, а на границе стыка фонем выполняется «сшивка» функций, представляющих собой зависимость амплитуды аудиосигнала от времени.

Конкатенационный метод синтеза речи дает хорошие результаты при аккуратной обработке и сшивке фонем. В частности, этот метод применяется в таких синтезаторах речи, как UTAU и Vocaloid (японские синтезаторы поющего голоса). В программе UTAU используются более простые алгоритмы обработки и синтеза, нежели в Vocaloid, поэтому и качество голоса, сгенерированного этим синтезатором, зачастую намного ниже. Однако при аккуратной работе композитора (тщательной ручной настройке параметров каждой фонемы) можно добиться хорошего качества звучания.

Тем не менее, нельзя не принимать во внимание некоторые существенные недостатки данного подхода к синтезу речи:

1. в моменты состыковки фонем возникают «артефакты» (на слух воспринимаемые как резкое изменение амплитуды и/или частоты звука), связанные с недостаточными требованиями на гладкость сшивки;

2. для изменения высоты тона и/или длительности фонемы, как правило, применяется преобразование Фурье (в составе алгоритма ресемплинга), которое тоже вносит свой вклад в образование звуковых дефектов.

Заметим, что данные дефекты генерируемого звукового сигнала возникают в основном из-за того, что мы работаем с временным его представлением. Их можно избежать, если работать с частотным представлением сигнала – проблема точного преобразования для изменения высоты тона или тембра звука при этом отпадает сама собой. В частности, в синусоидальном способе синтеза речи используется как раз частотное представление фонем. Основная идея данного подхода заключается в том, что речепроизводящий аппарат человека рассматривается как резонатор, в котором могут генерироваться колебания только некоторого ограничен-

ного набора частот [8]. Однако, в отличие от артикуляторного метода синтеза, при формантном методе моделируются не физиологические процессы образования речи, но результат этих процессов — акустические характеристики речевой волны [9–11].

Как уже было указано выше, в данной работе используется синусоидально-конкатенативный алгоритм синтеза, который позволяет получить для исходного сигнала аппроксимацию особого вида. Поставим задачу формально: пусть $\{F_k\}_{k=0}^{N}$ — отсчеты исходного дискретного сигнала.

Будем аппроксимировать его функцией следующего вида:

$$S(t) = \sum_{g=0}^{G} s_g(t) = \\ = \sum_{g=0}^{G} A_g(t) \cdot \sin\left(\omega_{g_0} \cdot (t - k_{g_0}) + O_g(t)\right),$$

где

$$A_g(t) = a_{g_0} + \sum_{p=1}^{N_{A_g}} a_{g_p} \cdot \sin\left(\gamma_{g_p} \cdot \left(t - \varphi_{g_p}\right)\right),$$

$$O_g(t) = \sum_{p=1}^{N_{O_g}} c_{g_p} \cdot \sin\left(\omega_{g_p} \cdot \left(t - k_{g_p}\right)\right).$$

Назовем функцию $s_g(t)$ g-й компонентой сигнала. Таким образом, задача алгоритма формантно–конструктивного синтеза — определение оптимальных характеристик компонент исходного сигнала:

$\left\{a_{g_p},\ \gamma_{g_p}, \varphi_{g_p}\right\}$ — амплитуды, частоты и фазы амплитуды g-й компоненты;

$\left\{c_{g_p}, \omega_{g_p}, k_{g_p}\right\}$ — амплитуды, частоты и фазы функций ее частоты.

Рассмотрим данный алгоритм более подробно. Как уже было указано, для анализа исходного речевого сигнала будем применять непрерывное вейвлет-преобразование. Однако заметим, что матрица данного преобразования избыточна — для определения параметров сигнала необходимо знать лишь расположение ее локальных экстремумов (так называемый скелетон). Поэтому вместо вычисления всех компонент матрицы организуем алгоритм трассировки скелетона матрицы $C_{x,y}$. В результате трассировки получим матрицу C, в которой определены только необходимые для последующего восстановления сигнала компоненты.

После трассировки аппроксимируем полученные временные зависимости частот и амплитуд найденных компонент сигнала. Для этого естественным решением будет повторно применить к ним алгоритм трассировки, что и было сделано в рамках данной работы. Однако стоит отметить один нюанс алгоритма — для определения низкочастотных составляющих сигнала период данных составляющих должен быть не больше удвоенного времени длительности исследуемого сигнала (в противном случае в матрицу его вейвлет–преобразования может не попасть ни одного экстремума производной). На практике же их период должен быть строго меньше длительности сигнала, так как по краям скейлограммы, как правило, присутствуют искажения (связанные с усечением анализирующего вейвлета в данных областях). У большинства компонент исследованных аудиосигналов были обнаружены подобные низкочастотные составляющие, для определения частот которых требуется другой метод — например, метод оптимизации (минимизации). Его суть заключается в аппроксимации сигнала синусом минимально возможной частоты, что сводится к задаче минимизации следующего функционала:

$$\Phi(\omega, \varphi_0, a, a_0) = \\ = \sum_{k=0}^{N} \left(P_k - \left(a_0 + a \cdot \sin\left(\omega k \cdot dt + \varphi_0\right)\right)\right)^2 \cdot dt,$$

где $\{P_k\}_{k=0}^{K}$ — отсчеты исследуемой компоненты сигнала, dt — шаг дискретизации.

Для минимизации данного многомерного функционала можно применять метод Ньютона — при хорошо выбранном начальном приближении он обеспечивает высокую точность и скорость сходимости. Однако практика показала, что для компонент речевого сигнала функционал Φ многоэкстремален, поэтому ошибка в оценке начального приближения может сильно сказаться на скорости сходимости (или и вовсе привести к расходимости метода). Поэтому применим метод сопряженных градиентов. При этом возникает проблема выбора метода одномерной минимизации (вдоль текущего направления). Хорошо себя зарекомендовал метод золотого сечения [12], но скорость его сходимости намного меньше, чем в методе Ньютона, который в нашем случае может давать не очень хорошие результаты даже в одномерном случае. Поэтому был разработан простой метод одномерной оптимизации, который показал бульшую устойчивость относительно начального приближения, чем метод Ньютона, при этом обладающий более высокой скоростью сходимости, чем метод золотого сечения.

По сути, это — модифицированный градиентный спуск: основная идея метода заключается в последовательном «движении» вдоль направления антиградиента функции. Начав из точки начального приближения, на каждом шаге вычисляется текущее значение антиградиента функции, на основе которого вычисляется величина шага в этом направлении. Если значение антиградиента мало или очень велико (по модулю), шаг следует брать относительно маленький, иначе его следует брать больше. Положим для определенности, что при равенстве модуля производной единице шаг алгоритма максимален. Исходя из этой идеи, конкретная зависимость величины шага от значения модуля производной функции (обозначим его как x) была взята,

из эвристических соображений, в следующем виде:

$$\delta(x) = h \cdot \frac{(\sqrt{x} - dh) \cdot e^{-\frac{x^2}{4(1-dh)}} + dh}{(1-dh) \cdot e^{-\frac{1}{4(1-dh)}} + dh}.$$

Множитель h задает начальную величину шага, при смене знака производной функции в рассматриваемой точке уменьшается в два раза, что обеспечивает сходимость алгоритма. Величина dh — это значение $\delta(x)$ на бесконечности (некоторое положительное число меньше единицы, позволяющее избежать проблемы «застревания» алгоритма на слишком крутых склонах исследуемой функции). Конкретный вид знаменателя был выбран с расчетом на то, чтобы функция $\delta(x)$ достигала своего наибольшего значения, равного 1, в единице.

Также, для ускорения сходимости, между тремя последовательными точками исследуемой зависимости проводится парабола, и если ее ветви оказываются направлены вверх, следующий шаг выбирается так, чтобы следующая точка соответствовала минимуму этой параболы.

Пример работы алгоритма: найдем минимум следующей функции:

$$f(x) = 1 - 2\frac{\operatorname{arctg}(x)}{\pi} + \sin\left(\frac{x}{2}\right)e^{\frac{-(x+10)^2}{8}} + \\ + \frac{x^2}{28(x+1)^2+1} + \frac{3(x-2)^2 \cdot \ln(x^2+0.1)}{2(x-2)^2+1},$$

стартовое значение: $x_0 = -10$.

Алгоритм сошелся к точке минимума $x \approx 0.0109039817198$ за 16 итераций.

Метод Ньютона при старте из точки x_0 либо находит максимум в окрестности -10, либо расходится (для модифицированного метода Ньютона, при неудачном выборе «поправок»). Встроенный метод поиска минимума функции в Wolfram Mathematica 10.2 при старте из x_0 сходится к точке $x \approx 0.0109039817168$ за 16 итераций.

Метод золотого сечения при старте из той же точки сошелся к точке $x \approx 0.0109039817178$ за 58 итераций.

2. СИНТЕЗ СИГНАЛА

Таким образом, проведя трассировку исходного аудиосигнала F в некотором диапазоне частот с последующей трассировкой полученных компонент (если нужно — определяя нижнюю их частоту описанным выше методом минимизации), в результате получим набор параметров

$$\left\{a_{g_p},\ \gamma_{g_p}, \varphi_{g_p};; c_{g_p}, \omega_{g_p}, k_{g_p}\right\}_{g=0}^{G},$$

который полностью определяет аппроксимацию функции $S(t)$.

Набор параметров аппроксимации для каждой фонемы удобно сохранять в базу данных, ставя им в соответствие некоторое текстовое представление (транскрипцию). Рассмотрим теперь алгоритм синтеза речи с использованием подобной базы данных. Пусть в качестве входных данных выступает текстовая строка, состоящая из разделенных пробелами транскрипций. Тогда синтез сведется к:

1. поиску каждой транскрипции текста в базе данных и получение соответствующего ей спектра;

2. конструктивному синтезу фонемы (расчету соответствующей функции аппроксимации) и выполнению сшивки на границе ее стыка с соседними.

В данной работе были исследованы аудиофайлы голосовой базы Zuiga Mizuki, некоторые из них (файлы гласных звуков) были аппроксимированы при помощи предложенного алгоритма синтеза. На рис. 1 представлены спектрограммы исходного и синтезированного сигналов для фонемы «и» (по горизонтальной шкале — частота в Гц, по вертикальной — Децибелы):

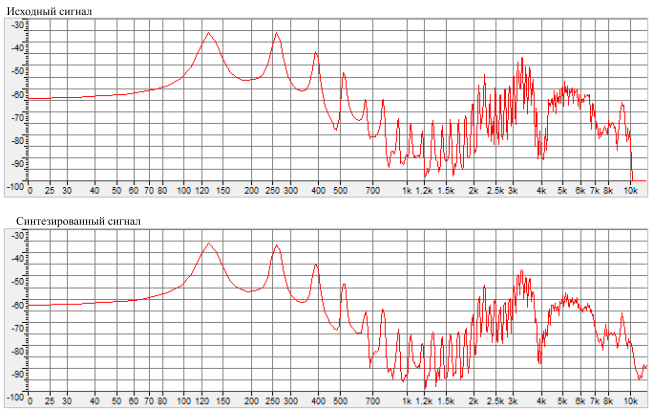

Рис. 1: Спектры исходного и синтезированного сигналов (фонема «и»)

Из графиков видно хорошее их согласование, хотя в высокочастотной области заметно их расхождение. Отметим также, что размер данных, потребовавшихся для восстановления сигнала, составил 29 Кбайт, что более чем в 3 раза меньше исходного объема информации (102 Кбайт).

ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены основы непрерывного вейвлет-преобразования и исследованы некоторые его свойства и особенности по сравнении с Фурье–преобразованием. На основе данного исследования, а также изучения основных подходов к задаче синтеза речи, были разработаны алгоритмы:

1. анализа аудиосигналов (трассировка скелетона матрицы его непрерывного вейвлет–преобразования);
2. аппроксимации полученных компонент сигнала (трассировка найденных компонент с применением алгоритма аккуратного исследования их низкочастотных составляющих).

Для реализации данных алгоритмов была написана программа на языке C++, также позволяющая проводить редактирование скелетона перед аппроксимацией и синтез сигнала по полученной аппроксимации. С ее помощью были исследованы фонемы, которые затем была синтезированы по меньшему объему данных.

[1] *Tzanetakis G., Essl G., Cook P.* Audio analysis using the discrete wavelet transform. Princeton NJ 08544 USA.
[2] *Chui Ch. K.* An Introduction to Wavelets: Department of Mathematics, Texas A&M University, College Station, Texas.
[3] *Павлов А. Н.* Методы анализа сложных сигналов: Учебное пособие для студентов физического факультета.
[4] *Пиуновский Е. В., Тропченко А. А.* Изв. Вузов. Приборостроение. 2012. **55**, № 3.
[5] *Сорокин В. Н.* Синтез речи. М.: Наука, 1992.
[6] *Рыбин С. В.* Синтез речи. Учебное пособие. СПб: Университет ИТМО, 2014.
[7] *Дремин И. М., Иванов О. В., Нечитайло В. А.* Вейвлеты и их использование. Физический институт им. П. Н. Лебедева РАН, 2001.
[8] *Лобанов Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи. Минск, «Белорусская Наука», 2008.
[9] *Klatt D. H.* JASA. 1987. **30**, № 3. P. 737.
[10] *Фант Г.* Акустическая теория речеобразования. М.: Наука, 1964.
[11] *Allen J. M., Sharon H. M., Klatt D. H.* From text to speech the MITalk system. Cambridge, MIT Press, 1987.
[12] *Калиткин Н. Н.* Численные методы. М.: Наука, 1978.

Wavelet analysis of audiosignals and speech synthesis

S. A. Nikonorov[a], A. N. Bogolubov

Faculty of Physics, Lomonosov Moscow State University. Moscow 119991, Russia
E-mail: [a]nics1543@gmail.com

The speech analysis and approximation algorithm based on the continuous wavelet transform was created. In the current work, audiofiles from the Zuiga Mizuki voice bank were analised and some of them (namely vowel sound files) were synthesised using the suggested algorithm.

PACS: 02.60.Gf.
Keywords: wavelet transform, speech audio analysis, speech synthesis algorithm, skeleton trace algorithm.
Received 25 June 2018.

Сведения об авторах

1. Боголюбов Александр Николаевич — доктор физ.-мат. наук, профессор, зав. отделением прикладной математики; e-mail: bogan7@yandex.ru.
2. Никоноров Сергей Алексеевич — студент магистратуры; e-mail: nics1543@gmail.com.