Сергей В. Дворянкин[1], Артём Е. Зенов[2], Роман А. Устинов[3], Никита С. Дворянкин[4]
[1,2]Московский государственный лингвистический университет,
ул. Остоженка, 38, стр. 1, Москва, 119034, Россия
[3]Финансовый университет при правительстве РФ,
Ленинградский пр-кт, 49, Москва, 125993, Россия
[4]Национальный исследовательский ядерный университет «МИФИ»,
Каширское ш., 31, Москва, 115409, Россия
[1]e-mail: DvoryankinS@linguanet.ru, http://orcid.org/0000-0001-6908-0676
[2]e-mail: zenov.a@linguanet.ru, http://orcid.org/0000-0002-9482-4945
[3]e-mail: public-ura@yandex.ru, http://orcid.org/0000-0002-8454-9951
[4]e-mail: nik.dvrn@gmail.com, http://orcid.org/0000-0002-1580-7179

КОДИРОВАНИЕ ИЗОБРАЖЕНИЙ СПЕКТРОГРАММ ДЛЯ ОБЕСПЕЧЕНИЯ ПЕРЕМЕННОЙ СКОРОСТИ ПЕРЕДАЧИ АУДИОДАННЫХ С СОХРАНЕНИЕМ КАЧЕСТВА ИХ ЗВУЧАНИЯ

DOI: http://dx.doi.org/10.26583/bit.2021.4.02

Аннотация. В приложениях, аудио контроля и фиксации в условиях информационно-технического противодействия, шумоподавления, формирования цифровых водяных знаков, звуковых отпечатков, защитных текстовых аудиомаркеров и др. требуется компактное представление речевых сигналов для последующей передачи-хранения при максимальном сохранении сходства звучания восстановленной речи с оригиналом, устранении сопутствующих шумов и помех. Предлагаемый аудиокодек основан на узкополосной синусоидальной Гауссовой модели анализа/синтеза речи, где ее представление в виде суперпозиции гармонических составляющих, взвешенных окном Гаусса, применимо для всех видов речевых фреймов, а также на универсальных и специальных методах построения и обработки изображений узкополосных динамических спектрограмм с применением к ним алгоритмов сжатия-восстановления, что позволяет регулировать скорость речевого потока в широких пределах *(1,2–16 Кбит/с)* с адаптацией к изменениям пропускной способности канала передачи-хранения аудиоданных, обусловленными как объективными факторами, так и действиями злоумышленника. Целью работы является выбор наилучших параметров на изображениях спектрограмм, которые уменьшают общий битрейт, устраняют влияние шумов и помех и позволяют посредством методов и алгоритмов спектральной инверсии восстановить речевой сигнал с прежним или лучшим качеством звучания. Параметры извлекаются из изображений спектрограмм, полученных с помощью кратковременного преобразования Фурье, используя методы выделения на спектральных срезах амплитуд, частот, фаз и треков развития отобранных опорных локальных и/или глобальных максимумов (пиков) речевого сигнала. В канал связи могут передаваться либо сами параметры, либо результаты сжатия-кодирования изображений для восстановления по ним на приемном конце образа исходной спектрограммы с выделением на ней параметров пиков с последующим синтезом по ним речи или для прямой спектральной инверсии восстановленного после сжатия изображения в речь. Возможна корректировка реконструируемой спектрограммы с использованием априорных сведений о речи диктора из заранее сформированной его голосовой базы данных.

Ключевые слова: защита речевой информации, аудио контроль, сжатие речи, улучшение речи, синусоидальная модель, инверсия спектрограмм, кратковременное преобразование Фурье, звуковой отпечаток.

Для цитирования: ДВОРЯНКИН, Сергей В. и др. КОДИРОВАНИЕ ИЗОБРАЖЕНИЙ СПЕКТРОГРАММ ДЛЯ ОБЕСПЕЧЕНИЯ ПЕРЕМЕННОЙ СКОРОСТИ ПЕРЕДАЧИ АУДИОДАННЫХ С СОХРАНЕНИЕМ КАЧЕСТВА ИХ ЗВУЧАНИЯ. Безопасность информационных технологий, [S.l.], т. 28, № 4, с. 22–38, 2021. ISSN 2074-7136. URL: https://bit.mephi.ru/index.php/bit/article/view/1376. DOI: http://dx.doi.org/10.26583/bit.2021.4.02.

Sergey V. Dvoryankin[1], Artem E. Zenov[2], Roman A. Ustinov[3], Nikita S. Dvoryankin[4]
[1,2]Moscow State Linguistic University,
Ostojenka str., 38/1, Moscow, 119034, Russia
[3]Financial University under Government of the Russian Federation
Leningradsky Prospekt, 49, Moscow, 125993, Russia
[4]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe shosse, 31, Moscow, 115409, Russia
[1]e-mail: S.Dvoryankin@linguanet.ru, http://orcid.org/0000-0001-6908-0676
[2]e-mail: zenov.a@linguanet.ru, http://orcid.org/0000-0002-9482-4945
[3]e-mail: public-ura@yandex.ru, http://orcid.org/ 0000-0002-8454-9951
[4]e-mail: nik.dvrn@gmail.com, http://orcid.org/0000-0002-1580-7179

Spectrogram image encoding to provide variable audio data rates and preserve its sound quality

DOI: http://dx.doi.org/10.26583/bit.2021.4.02

Abstract. In the applications of audio control and fixation in the conditions of information-technical counteraction, noise clearing, formation of digital watermarks, audio fingerprinting, protective text audio markers, etc., a compact representation of speech signals for subsequent transmission-storage is required while maximal preserving the similarity of the sound quality of restored speech with the original, elimination of accompanying interferences. The proposed audio codec is based on the narrow-band sine Gaussian model of speech analysis/synthesis, where its representation as a superposition of harmonic components weighted by a Gaussian window applies to all types of speech frames, as well as on universal and special methods of construction and image processing of narrow-band dynamic spectrograms, in particular, by the application of compression-recovery algorithms to them, which will allow to regulate the speech stream speed within a wide range of 1.2–16Kbit/s with adaptation to changes of the audio data transmission-storage channel bandwidth, caused, in particular, by both objective factors and the actions of an intruder. This work aims to select the best parameters on the spectrogram images that reduce the overall bitrate, remove the influence of noise and interference and allow using of spectral inversion methods and algorithms to recover the speech signal with the same or better quality. The parameters are extracted from the spectrogram images obtained using of the short-time Fourier transform, using methods to extract the amplitudes, frequencies, phases and development tracks of selected local or global maxima (peaks) of the speech signal on the spectral slices. The communication channel can transmit either the parameters themselves, or the results of compression-encoding of the image to restore the image of the original spectrogram with the selection of peak parameters already on it with the subsequent synthesis of speech or for direct spectral inversion of the image into speech. It is possible to correct the reconstructed spectrogram by using a priori information about the speaker's speech from his pre-generated voice database.

Keywords: speech information protection, audio control, speech compression, speech enhancement, sinusoidal model, spectral inversion, short-term Fourier transform, audio fingerprinting.
For citation: DVORYANKIN, Sergey V. et al. Spectrogram image encoding to provide variable audio data rates and preserve its sound quality. IT Security (Russia), [S.l.], v. 28, n. 4, p. 22–38, 2021. ISSN 2074-7136. URL: https://bit.mephi.ru/index.php/bit/article/view/1376. DOI: http://dx.doi.org/10.26583/bit.2021.4.02.

Введение

Кодирование речи имеет дело с проблемой получения компактного, сжатого представления речевых сигналов (РС) для эффективного цифрового хранения или передачи. Методы кодирования речи используются для улучшения использования полосы пропускания и энергоэффективности в таких приложениях, как цифровая телефония, передача мультимедийного контента, компьютерная стеганография, голосовое управление и, конечно же, безопасность связи в каналах коммуникаций, которые используют РС в цифровом формате для облегчения его обработки, хранения и передачи.

Особенно остро проблема речевого сжатия-восстановления встает в приложениях беспроводной и спутниковой связи, где пропускная способность в ряде случаев ограничена, в системах компьютерной стеганографии, для защиты авторских прав с ограничениями на объем аудиоданных, используемых в качестве цифровых водяных знаков, а также в системах аудио контроля и фиксации в условиях внешних мешающих факторов, когда требуется адаптивная каналу связи регулировка скорости передачи-хранения, необходимая для представления РС с минимальным количеством битов, с сохранением при этом его качества звучания (разборчивость, узнаваемость). Поэтому снижение скорости передачи аудиоданных за счет кодирования параметров новых описаний РС с нейтрализацией влияния присутствующих шумов и помех до сих пор актуально.

Для того чтобы уменьшить скорость передачи речевого сигнала используются разные методы удаления избыточной информации из речевого сигнала. Существует следующее разделение методов кодирования речи на основные группы [1, 2], условно различаемые по типу кодирования и скорости передачи аудиоданных:

– основанные на описании и предсказании волновой формы РС, обеспечивая при этом скорость кодирования *16–32 Кбит/с*;

– основанные на ортогональных преобразованиях РС с выделением и обработкой параметров с низкой скоростью их передачи в диапазоне *1,2–4,8 Кбит/с*;

– смешанного типа со средней скоростью передачи аудиоданных *4,8–16 Кбит/с;*

– речеэлементное (фонетическое и лингвистическое) кодирование (*0,1–1,2 Кбит/с)*;

Работы по последнему типу пока еще не доведены до практического применения, хотя скоро это должно случиться.

Как правило, в речевой связи используются аудиокодеки с фиксированными скоростями речепреобразования в широком диапазоне *1,2–64 Кбит/с,* в то время как на практике часто требуется плавная регулировка скорости кодирования, адаптивная к изменениям пропускной способности речевого канала передачи-хранения (связи), на который может быть оказано деструктивное ухудшающее воздействие.

Описанный в данной статье метод относится ко второй и третьей группам. Возможно применение некоторых подходов из этого метода в кодеках 4-й группы. Использование метода на высоких скоростях кодирования 1-й группы также возможно, но нецелесообразно из-за высокой вычислительной нагрузки кодера. Здесь достаточно применения традиционных, «простых» кодеков волновой формы РС.

В предлагаемой технологии образного кодирования речи используются либо результаты обработки спектрограммы, с выделением на ней параметров синусоидально Гауссовой модели РС, либо регулируемое по глубине сжатие-восстановление изображений спектрограмм – с последующим синтезом по ним речи и с обеспечением требуемой пропускной способности в широком диапазоне скоростей передачи с сохранением качества.

В первом случае из последовательных выборок кратковременного спектрального анализа извлекаются параметры модели, которые затем в кодере преобразуются в двоичные биты. После этого двоичный сигнал передается декодеру. На этапе синтеза декодер восстанавливает из полученных двоичных битов опорные синусоиды и суммирует их, одновременно взвешивая и сглаживая временным окном Гаусса.

Во втором, для сжатия-восстановления речи используются алгоритмы сжатия-восстановления изображений полутоновых и построенных на их основе бинарных узкополосных спектрограмм. Несмотря на то, что образное спектральное описание РС в битах может значительно превышать его волновое, тем не менее алгоритмы сжатия

изображений существенно уменьшают его объем для реализации требуемых скоростей передачи речи с высоким качеством.

В результате постобработки в декодере, используя восстановленные и реконструированные спектральные срезы, можно восстановить речевой сигнал по звучанию по возможности близким к исходному посредством алгоритмов спектральной инверсии Маккалая-Квартьери, Гриффина-Лима и др. [3–5]. Звучание переданного и принятого РС будет тем более похожим, чем ближе друг к другу спектрограммы обоих сигналов, определяющие фонетическую функцию Пирогова, отвечающую за смысловое содержание РС.

Схема данной работы организована следующим образом: в следующем разделе представлена реализация усовершенствованной синусоидальной модели анализа-синтеза РС, за этим следует обсуждение системы образного анализа-синтеза речи, стратегий выделения и кодирования параметров модели РС и результатов экспериментов.

1. Уточненная синусоидальная Гауссовская модель анализа-синтеза речи

В работе [3], как результат вокодерной концепции, предложенной для разработки новой методики анализа-синтеза речи, была представлена синусоидальная модель РС, характеризующаяся амплитудами, частотами и фазами составляющих речевые фреймы синусоид:

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \varphi_l) \quad (1)$$

где n – номер временного отсчета; L – количество опорных значимых синусоид в фрейме; A_l – амплитуда; ω_l – частота; φ_l – начальная фаза каждой l-й синусоиды, составляющей звучание этого речевого фрейма. В последующих исследованиях было показано, что эта модель может быть широко применима в приложениях речевой обработки и обеспечивает достаточно высокое качество восстановленной речи даже при относительно низкой скорости передачи данных [1].

Дальнейшее развитие этой модели [5, 6] показало, что РС при размере, анализируемого K-го кадра речи *20–40 мс* с шагом анализа *4–12 мс* скользящим временным усеченным окном Гаусса с длительностью выборки в $N=1024$ отсчетов [5] может быть описан:

$$s(n) = \sum_{l=1}^{L} A_l e^{\frac{-n^2}{2\sigma}} \cos(\omega_l n + \theta_l(n) + \varphi_l) + e(n) \quad (2)$$

где помимо понятных параметров из (1) ещё добавились: θ_l – функция нелинейности фазы, σ – эффективная ширина окна Гаусса, $e(n)$ – остаточный сигнал (шум).

В виде (2) исходный речевой сигнал можно рассматривать как суперпозицию узкополосных сигналов, вейвлетов Морле, или коротких синусоид, взвешенных окном Гаусса. Такое представление (2) можно распространить и на другие акустические сигналы.

Заметим, что при минимизации длительности временного шага анализа формула *(2)* сводится к описанной ранее синусоидальной модели Маккалая-Квартьери (1).

Указанные выше параметры узкополосных составляющих речи (2) могут быть оценены на этапе анализа по результатам кратковременного преобразования Фурье (КПФ) с базой быстрого преобразования Фурье (БПФ) – N. Как правило $N=1024$.

Из свойств преобразования Фурье известно, что умножение функции на синусоиду во временной области приводит к сдвигу спектра этой функции на частоту этой синусоиды в частотной области. Для функции временного окна Гаусса спектром тоже является функция колокола Гаусса, но уже другой ширины, смещенная на частоту

синусоиды модели (2). Следовательно вся информация о работе синусоидальной Гауссовой речевой модели (2) будет содержаться в амплитудах, частотах и фазах локальных максимумов (ЛМ) или пиках мгновенного спектра, а также треках, контурах их развития на изображениях спектрограмм.

Волновая форма РС (осциллограмма) и соответствующее ей изображение узкополосной динамической полутоновой спектрограммы показаны рис. 1.

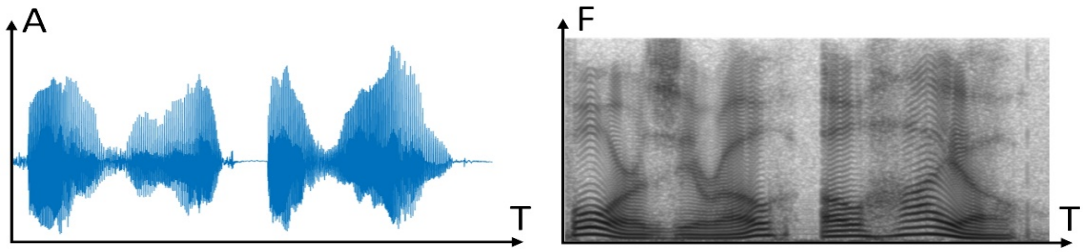
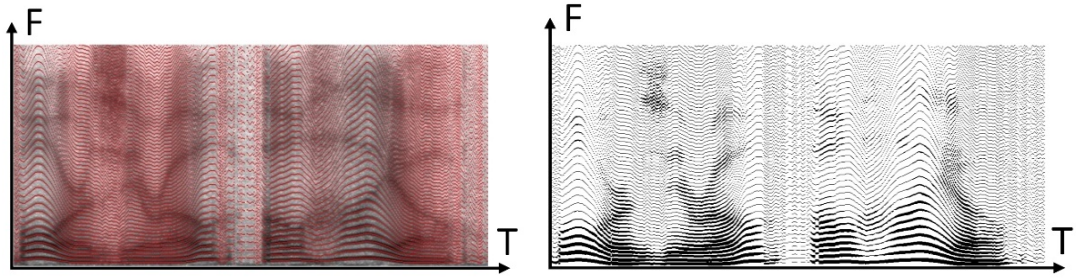


Рис. 1. Изображение осциллограммы (а) и полутоновой (в уровнях серого) спектрограммы фрагмента речевого сигнала (б).
Fig. 1. Image of the oscillogram (a) and grayscale spectrogram of the speech signal fragment (b).

Полутоновая спектрограмма фрагмента исходной речи (рис. 2,а) с выделенными на ней красным цветом треками ЛМ или пиков составляющих её синусоид и построенный на их основе её бинарный образ (рис. 2,б) представлены на рис. 2.



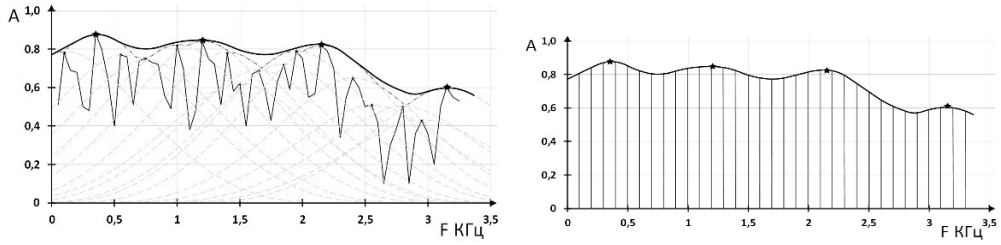
а) *б)*

Рис. 2. Полутоновая спектрограмма с выделенными красным треками пиков (а) и построенная по ним по правилу «амплитуда пика равна толщине трека» бинарная спектрограмма (б).
Fig. 2. Half-tone spectrogram with the peak tracks highlighted in red (a) and the binary spectrogram based on the rule "peak amplitude equals track thickness" (b).

На полутоновых спектрограммах рис. 1, 2 и далее, в координатах время (ось абсцисс) и частота (ось ординат) уровнем серого цвета указаны мощность РС в конкретном узле сетки его частотно-временного описания. Самый мощный участок – черного цвета, самый слабый – белого.

Схематичное изображение мгновенного спектра, спектрального среза (столбца) спектрограммы, с пиками опорных синусоидальных составляющих приведено на рис. 3,а.

Отметим, что пики спектрограммы, устойчивы к шумам и искажениям, извлекаются и используются в качестве характерных признаков речевого фрагмента, в том числе и в качестве его звукового отпечатка. Точка в спектрограмме считается пиком, если ее амплитуда выше, чем у соседних точек.



а) *б)*

Рис. 3. Схема спектрального среза (столбца) полутоновой спектрограммы с пиками элементарных синусоид (черные точки), результатами их Гауссового сглаживания (звездочки) (а) и с выделением спектральной огибающей среза (б).

Fig. 3. Schematic of the spectral slice (column) of the half-tone spectrogram with the peaks of elementary sinusoids (black dots), the results of their Gaussian smoothing (asterisks) a) and with separation of the spectral envelope of the slice (b).

Одномерное Гауссово сглаживание используется для подавления мало значимых и «ложных» локальных максимумов (ЛМ) в векторе, который соответствует срезу (столбцу) в спектрограмме. Его результат проиллюстрирован на рис. 3,а штрихпунктирной линией.

Здесь вначале выделяются все ЛМ среза выше заданного порога (черные точки). Затем на каждый локальный максимум накладывается колокол Гаусса (все штрих линии). А точечные максимумы суммы всех наложенных Гауссианов в виде спектральной огибающей (штрихпунктирная линия) являются результатом этого первого сглаживания.

В примере на рис. 3,б, после первого сглаживания, *11* мало значимых по вкладу в общее звучание анализируемого отрезка РС максимумов оказались подавлены, а общее количество пиков (точки на рис. 3,а) уменьшилось с *17* до *6*. Дополнительное повторное Гауссово сглаживание еще больше уменьшает их количество – до *4*-х (звездочки).

Таким образом, сначала к конкретному столбцу спектра применяется Гауссово сглаживание. Затем определяются все оставшиеся опорные локальные максимумы (ЛМ) и извлекаются их параметры для синусоидальной Гауссовой модели. Только те максимумы, которые превышают заданное значение порога, выбираются в качестве рельефных значимых пиков столбца.

Описанная процедура Гауссового сглаживания применяется начиная с первого столбца (среза) спектрограммы и до её последнего столбца (среза).

Дополнительное сглаживание применяется посрезно для отбора глобальных максимумов (ГМ), пригодных для кодирования спектральной огибающей при реализации низких скоростей передачи РС и/или для поиска эталонного образца РС в базе данных голоса диктора по так сформированному звуковому отпечатку.

После определения глобальных пиков сложная спектрограмма преобразуется в компактную последовательность координат ГМ и временных расстояний между ними. Такой список координат, называемый хэшем РС или «картой созвездий», поскольку глобальные пики на спектрограмме выглядят как множество звезд на небе, может быть применен для поиска эталонных образов речи из базы данных голоса диктора, если исходная речь была искажена шумами и помехами и требуется повысить её качество.

Такая технология использования звуковых отпечатков (audio fingerprinting) для шумоподавления РС получает все большее развитие и распространение [7], может быть применена и в нашем случае одновременно с переменным кодированием речи.

2. Система образного анализа-синтеза речи

Система анализа-синтеза речи, соответствующая рассмотренной узкополосной синусоидальной Гауссовой модели РС, с учётом свойств органов слуха, эффектов частотного и временного маскирования, психоакустики и др., дополненная блоками обработки и контурного анализа узкополосных спектрограмм [8] представлена на рис. 4. За основу взята схема системы из работы [5].

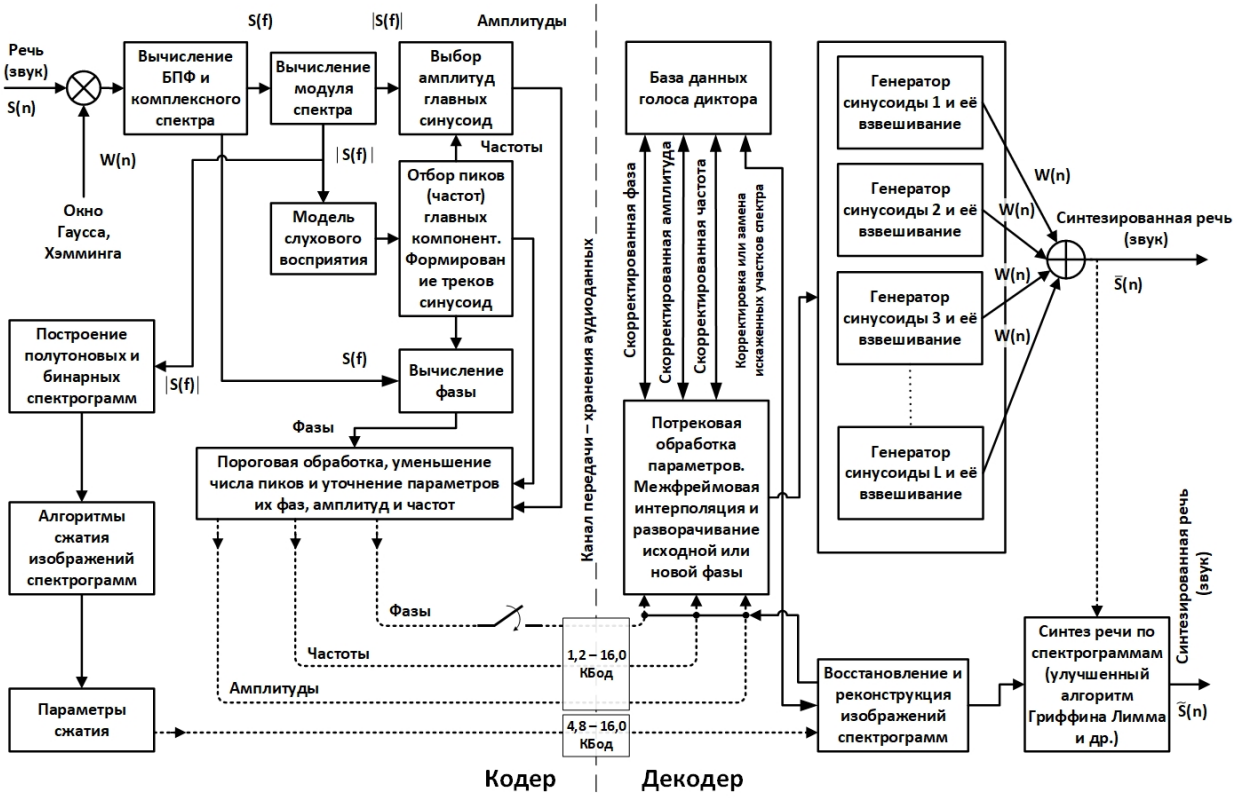


Рис. 4. Система образного анализа-синтеза речевого сигнала на основе синусоидальной узкополосной модели, свойств слуха и обработки изображений спектрограмм: а) анализирующая часть – кодер; б) синтезирующая часть – декодер.

Fig. 4. System of figurative analysis-synthesis of a speech signal on the basis of a sinusoidal narrowband model, properties of hearing and processing of spectrogram images: a) analyzing part – coder; b) synthesizing part – decoder.

В предлагаемой системе заложены две техники компрессии речи с сохранением её качества в широком диапазоне изменения скоростей передачи канала связи.

В первой – блок анализа (кодер) используется для извлечения параметров пиков из спектральных срезов для каждого временного речевого кадра или кадров.

Процедура последовательного вычисления спектральных срезов состоит в разделении протяженного по времени речевого сигнала на короткие, перекрывающиеся отрезки (кадры), с последующим взвешиванием их в окне и вычислением через быстрое преобразование Фурье (БПФ) комплексного, а из него и амплитудного и фазового спектров поочередно на каждом отрезке.

На каждом шаге анализа, после преобразования речевого кадра в частотную область с помощью описанного кратковременного преобразования Фурье (КПФ), все пики с соответствующими им параметрами извлекаются из спектра с использованием одного из

алгоритмов пикового подбора значений амплитуд частот и фаз, составляющих речь элементарных синусоид [2, 3, 5], о чем подробнее будет сказано ниже.

Извлеченные параметры кодируются, передаются, декодируются и используются в блоке синтеза для восстановления новой речи в гребенке синусоидальных генераторов, выход каждого звена которой модулируется амплитудой на частоте пика с исходной или искусственной фазой и, добавляясь к другим синусоидальным волнам, взвешивается временным окном Гаусса. Заметим, что эту операцию можно проводить и с помощью процедуры обратного БПФ, что может дать серьезный выигрыш по времени обработки.

На этом этапе синтеза учет изменений в спектральных компонентах треков ЛМ узкополосных составляющих отслеживается с использованием концептуальных понятий «рождение» и «смерть» синусоиды, лежащих в основе принятого синусоидального представления [2, 3].

По второй технике компрессии – ансамбль получаемых спектрально-временных разверток модуля спектра, рассматривается, как изображение спектрограммы с возможным применением к нему всего арсенала средств цифровой обработки изображений [5, 6, 8] для решения той или иной задачи. Как, например, в нашем случае сжатие-восстановление изображения спектрограммы с одновременным удалением из спектральных описаний следов шумов и помех, и передачей параметров сжатия по каналу связи с последующим синтезом в декодере по восстановленной спектрограмме нового РС с требуемыми свойствами.

Синтез речи или звука по изображению реконструированной спектрограммы может проходить с использованием одного из алгоритмов спектральной инверсии. Наиболее известным, но требующим множества итераций, считается алгоритм Гриффина-Лима [9].

При наличии шума и помех производится удаление или сглаживание следов их пиковых треков в кодере в случае использования первой техники сжатия речи, или корректировка изображения принятой восстановленной спектрограммы, подлежащей синтезу, во втором случае. Здесь возможно использование заранее сформированной голосовой базы диктора для корректировки или даже замены искаженных участков изображения спектрограмм с поиском эталонов по совпадению «звездных карт».

3. Методы извлечения, сокращения и кодирования пиковых параметров

Важнейшей частью в системе образного анализа-синтеза является обнаружение пиков, поскольку речь восстанавливается в декодере по соответствующим им параметрам.

Рассмотрим методы уменьшения числа параметров, битов их квантования, для адаптации требуемой скорости передачи к пропускной способности (ПС) канала передачи-хранения и улучшения качества восстановленного сигнала. Здесь возможны к применению различные стратегии на основе результатов [2, 3, 5, 8].

Выделение локальных максимумов (пиков)

Почти во всех системах анализа синусоидальных сигналов обнаружение пиков и оценка параметров выполняются в частотной области [2, 3], поскольку каждая стабильная синусоида сдвигает вершину образа колокола Гаусса на свою частоту. Для того чтобы считать речевой сигнал квазистационарным, длина каждого кадра анализа-синтеза на основе КПФ должна быть достаточно мала. В предлагаемой системе рис. 4 кодер делит оцифрованный с частотой *8 КГц* речевой сигнал на основные кадры от *20* до *40 мс*, с шагом анализа (перекрытия) *4–12 мс*, а затем преобразует их в частотную область с помощью быстрого преобразования Фурье (БПФ).

Чтобы отличить спектрально близкие синусоиды друг от друга на этапе анализа используется скользящее усеченное окно Гаусса, поскольку оно имеет очень хорошую

структуру и низкий уровень боковых лепестков [10], что практически не дает ложных ЛМ и улучшает качество восстановленной речи после синтеза по опорным параметрам.

Заметим, что снижение в принятой модели РС числа синусоидальных компонент с *N/2* (*N* база КПФ на основе алгоритма БПФ, как правило, выбирается равной *1024*) до *L* (максимальное число обычно выбирается равным *80* [2] или *64* [5]) не должно сильно сказываться на качественных и смысловых характеристиках звука и речи, синтезируемых по амплитудно-частотно-фазовым параметрам пикового подбора треков узкополосных составляющих РС на изображениях спектрограмм.

Среднее количество пиков, входящих в каждый основной кадр, согласно [2, 5] составляет *15–30*, максимальное *64–80*. Дальнейшее сокращение числа синусоидальных компонент возможно за счет использования свойств слухового восприятия и анализа изображений спектрограмм [11].

Простейший метод извлечения из спектра синусоидальных волн речевого сигнала заключается в выборе на спектральном срезе всех локальных максимумов с их амплитудами, частотами и фазами, где пик указывает на наличие взвешенной окном Гаусса синусоидальной волны. Этот метод аудио кодирования, является очень простым и эффективным, чтобы обеспечить среднюю и высокую скорость передачи данных. Однако для достижения низкой скорости передачи необходимо выбрать небольшое количество опорных, значимых синусоид. Поэтому естественным продолжением приведённого метода является применение порога обнаружения, когда все локальные максимумы выше порога интерпретируются как пики составляющих речь главных синусоид.

В наиболее распространённой процедуре [2] для каждого полученного спектрального среза (столбца изображения спектрограммы) пики выбираются путем нахождения места изменения спектрального наклона от положительного к отрицательному (черные точки ЛМ на рис. 3). Более точная техника использует параболу, которая подгоняется к пику, и местоположение ее вершины кодируется как частота пика [11].

Уменьшение числа пиков, фаз и задание порога

Следующая процедура определяет выбор лучших, опорных *L* синусоидальных волн в каждом речевом кадре спектрограммы, состоящим из группы спектральных срезов, вычисляемых с заданным шагом анализа. Необходимо также учитывать, что выбираемое значение *L* также зависит от требуемой скорости передачи-хранения данных. Процедура состоит из следующих шагов:

1. Для каждого кадра после преобразования его в частотную область, происходит отбор всех пиков, расположенных на всех спектральных срезах (столбцах), составляющих кадр. В зависимости от выбранного шага анализа в один кадр могут входить *2–10* срезов.

2. Если группы пиков в суммарном столбце достаточно близки друг к другу, то выбирается и остается самый большой пик для их совместного представления в кадре.

Фазы для отобранных пиков оцениваются с помощью процедуры извлечения фаз, предложенной Mcaulay и Quatieri в [3]. После выполнения этой процедуры количество фаз уменьшается с незначительным влиянием на качество речи, поскольку человеческое ухо менее чувствительно к фазовым искажениям, поэтому их устранение оправдано.

Для низкоскоростного кодирования можно совсем отказаться от вычисления фаз в кодере и синтезировать искусственную фазу (разницу фаз) в декодере в соответствии с зависимостью значений фазы синусоиды от развития изменений её трека на спектрограмме.

Также может быть использована техника задания порога [2, 5], которая считается наиболее эффективной в том смысле, что она уменьшает количество пиков без ущерба для восприятия голоса с учетом психоакустики и свойств слуха. Пороговое значение

выбирается так, что все пики ниже него будут устранены. При этом уменьшается не только количество амплитуд ЛМ, но и соответствующее им количество частотных мест и соответствующих фаз. Таким образом, пороговый метод уменьшает общую скорость передачи данных и улучшает восстановленную речь, путем фильтрации пиков присутствующего шумового сигнала, амплитуда которых меньше порогового значения.

С другой стороны, увеличение порога выше определенного значения приводит к повреждению речевого кадра из-за фильтрации важных информационные пиков. Поэтому пороговое значение должно быть выбрано с учетом этого.

Как развитие рассмотренных процедур отбора опорных пиков стоит рассматривать медианное сглаживание изображения спектрограммы и особенно Гауссовское сглаживание (рис. 3). В последнем случае в пределе можно ограничиться треками движения *3÷4* глобальных максимумов, как правило, связанных с движением формант (рис. 3,б).

Техника кодирования фаз, частот и амплитуд опорных пиков

Биты, используемые для квантования фаз, могут быть уменьшены путем минимизации их энтропии [2]. Тогда в кодере будет предсказываться разница между текущей фазой и ее прошлым значением и кодироваться будет разность фаз, а не сама фаза.

Для кодирования разницы фаз будет достаточно *4 бит* [2, 3].

Минимальное количество битов, необходимое для кодирования каждого частотного местоположения ЛМ при ширине спектра *N/2=512*, составляет *9 бит*. С помощью *6 бит* можно и удобней представлять разницу местоположений между соседними пиками на суммарном спектральном срезе.

Известно, что высокочастотные компоненты оказывают меньшее влияние на восприятие речи. Поэтому более высокочастотные позиции ЛМ могут быть квантованы с использованием меньшего количества битов чем более низкочастотные. Это позволяет несколько снизить скорость передачи данных, сохраняя при этом качество речи практически на прежнем уровне. Для реализации этой идеи можно перейти от линейного представления спектрограммы к ее мэл-кепстральному виду. Возможны и другие решения. Среднее число бит на частотную позицию пика и в этом случае также можно принять равным *6 бит*.

Техника кодирования синусоидальной амплитуды также важна, поскольку амплитуда пика подвержена изменениям в процессе его трекового развития на спектрограмме. Предположим, что есть амплитуды ЛМ на суммарном срезе кадра:

$$x(l) = [amp1, amp2,...,ampL],$$

где L – количество рассматриваемых пиков. Тогда предложенная процедура кодирования сводится к взятию $Log_2(x(l))$ амплитуды, чтобы уменьшить динамический диапазон. На это также можно отводить по *6 бит* для каждого оставленного пика.

4. Оценка скорости передачи-хранения аудиоданных

После выполнения методов редукции и кодирования имеем: L амплитуд пиков плюс L их частотных позиций плюс ($0,5L$) значений фаз для каждого основного кадра. В [2, 3] отмечается целесообразность оставления *6 бит* для каждой амплитуды и частоты ЛМ и *4 бит* для каждой его фазы.

Таким образом, оцениваемая скорость передачи аудиоданных для каждого кадра равна – $(6L + 6L + 4(0,5L)) = 14L$ *бит/кадр*.

Общая скорость передачи данных R может быть вычислена как:

$$R = 14L(бит/кадр) * K(кадр/с) = 14LK\ бит/с.$$

В случае же использования искусственно синтезированной фазы, не требующей бит для своего кодирования, скорость может составить:

$$R = 12L(\text{бит/кадр}) * K(\text{кадр/с}) = 12LK \text{ бит/с.}$$

Некоторые дополнительные биты также могут быть использованы для управления, обнаружения и исправления ошибок.

Можно подсчитать, что минимально достижимая скорость передачи параметров речи предложенным способом параметризации пиков составит *1,2 Кбод* для *L=4, K=25* при длительности кадра *40 мс.*

5. Особенности восстановления речи по полученным параметрам и описаниям

Как уже отмечалось используются сразу две техники параметрического сжатия-восстановления речи.

В первом случае декодер используется для восстановления исходного сигнала путем декодирования параметров пиков, извлеченных на этапе кодирования, как показано на рис. 4. Эти параметры затем используются для восстановления кадров речи путем линейного суммирования синусоидальных волн различных амплитуд, частот и фаз, взвешенных окном Гаусса.

Во втором случае – в приёмнике полученное двоичное представление параметров сжатого изображения исходной спектрограммы РС преобразуется в реконструированную форму, при этом восстановленный образ спектрограммы должен быть максимально близок к исходному, и уже по нему проводится синтез нового РС по извлечённым параметрам пиков или сразу синтез по спектрограмме алгоритмом спектральной инверсии [12].

Заметим, что полученная в декодере синтетическая форма волны РС хорошо совпадает с исходной волной в случае использования оригинальной фазы треков, что правда требует большего числа бит на её кодировку, и/или сохраняет общую форму частотной огибающей исходного сигнала для искусственно подобранной фазы, как для первой, так и второй техники сжатия. Качество звучания (разборчивость, узнаваемость) синтезированных РС для обоих типов выбираемой фазы для высоких и средних скоростей передачи параметрических данных практически не отличается от оригинальной речи, звука.

На этапе синтеза возможна корректировка изображений спектрограмм на основе априорных сведений о параметрах голоса диктора и алгоритмов машинного обучения (рис. 4), с последующей отрисовкой на восстанавливаемых спектрограммах новых треков ЛМ и синтезом по ним новых речевых кадров. В процессе «глубокой» корректировки возможна также замена «испорченных» шумом и помехами участков спектра на эталонные из базы данных голоса конкретного диктора, найденных по их звуковым отпечаткам (звёздным картам).

6. Экспериментальные результаты

В проведённых экспериментах на базе системы образного анализа-синтеза (рис. 4) размер эффективной ширины временного окна выбирался равным *2,5*-кратному среднему периоду основного тона. Использование усечённого до базы БПФ (*N=1024)* окна Гаусса и/или окна Хэмминга, дополненного до базы *N* нулями, в принципе показало схожие результаты. Размер основного кадра находился в пределах *20–40 мс*. Тестировались две техники передачи-хранения РС с переменной скоростью

Техника передачи синусоидальных параметров речи

Уменьшение общего числа пиков и выбор порогового значения проводились с учетом рассмотренных ранее стратегий и определялись следующим образом, несколько отличным от [5]:

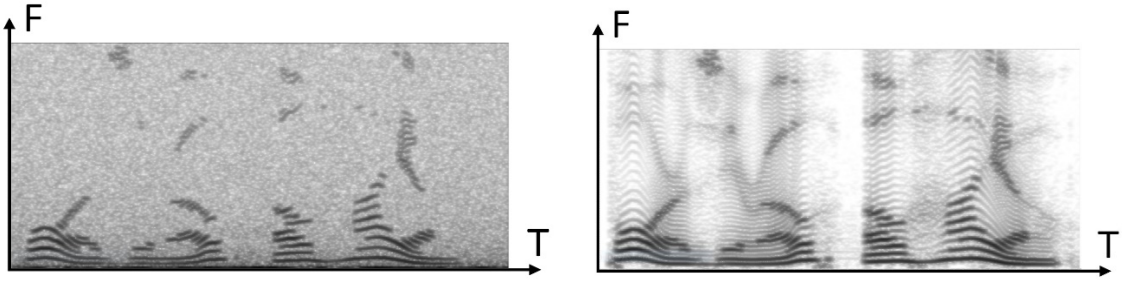
– совместная мощность оставленных максимальных пиков опорных синусоид составляла не менее 80% (граница спектрального спада) общей мощности всех пиков исходного спектрального среза;

– отбираемые главные синусоиды размещались на частотно-временной сетке синтезируемой спектрограммы в зоне присутствия как минимум двух разных формант или ГМ на исходной спектрограмме.

В результате моделирования системы образного анализа-синтеза РС (рис. 4) и экспериментов на ней было установлено, что минимальное число L таких главных синусоидальных компонент, выявляемых на изображениях узкополосных спектрограмм, может составлять от *8*-ми до *16*-ти как на вокализованных, так и на невокализованных участках речи.

Скорость передачи данных по предлагаемой технике может варьироваться от *1,2 Кбит/с* до *16 Кбит/с* или фиксироваться на конкретном значении из данного диапазона. Восстановленная в декодере речь была полностью разборчива.

Для улучшения комфортности звучания так синтезируемой речи на низких (ниже *4,8 Кбод*) скоростях рекомендуется добавлять (подмешивать) к ней фоновый коричневый шум, а главные гармоники в области верхних частот (формант) подчеркивать усилением их громкости функцией эквалайзера (рис. 5,а). Также, здесь, перед синтезом речи по восстановленной спектрограмме рекомендуется делать временную протяжку пиковых треков отобранных опорных синусоид, формируя следы их присутствия как на исходной спектрограмме, но с единичной амплитудой ЛМ (рис. 5,б). Это улучшает естественность и узнаваемость синтезированной речи.



а) *б)*

Рис. 5. Спектрограммы речи, синтезированной по параметрам главных пиков L≤8 с добавлением комфортного шума (а) и временной протяжкой треков главных синусоид (б).
Fig. 5. Speech spectrograms synthesized using the parameters of the main peaks L≤8 with the addition of the comfort noise (a) and the time stretching of the tracks of the main sinusoids (b).

Дополнительно уменьшить битрейт, как уже отмечалось, можно за счет медианного и Гауссовского сглаживания группы срезов спектрограммы (рис. 3), составляющих кадр, учитывая и оставляя только глобальные максимумы (ГМ) с $L=3\div4$, связанные с местоположением формант в низкочастотной области речевого спектра. Затем используя логарифмирование значений каждой амплитуды и частотной позиции ГМ можно выделять по *5 бит* на каждый из этих параметров, обеспечивая скорость передачи в районе *0,8–1,2 Кбод* с сохранением хорошей разборчивости. Результат такого синтеза по параметрам *3÷4*-х формант и жестко заданной частотой основного тона в *125 Гц* (рис. 3,б) показан на рис. 6. Еще больше скорость передачи можно снизить в случае разметки РС на паузные и непаузные участки с кодированием только последних и используя методы машинного обучения и априорные сведения о параметрах голоса диктора, извлекаемых из заранее сформированной базы данных его голоса.

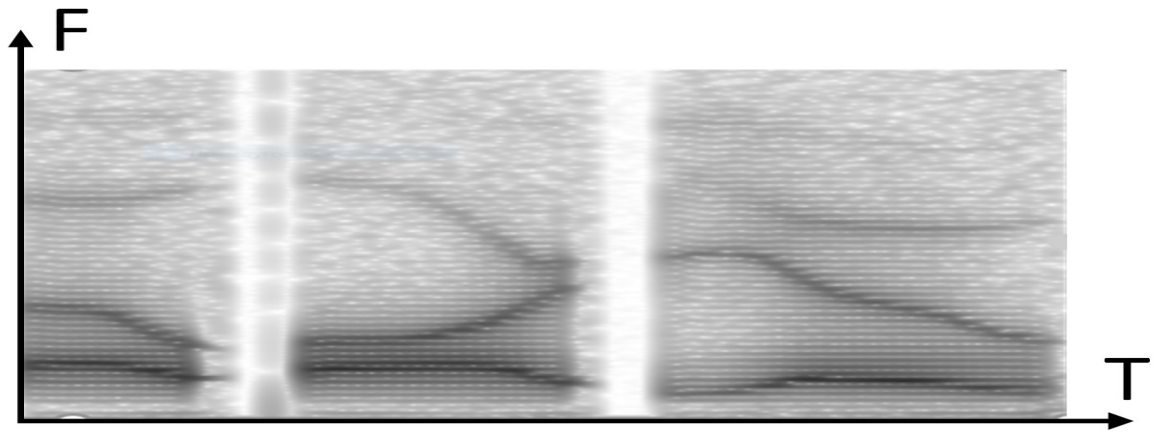


Рис. 6. Спектрограмма искуственной речи, синтезированной по трекам 4-х формант и трекам искусственно заданных гармоник с частотой основного тона 125 Гц.
Fig. 6. Spectrogram of the artificial speech synthesized from the tracks of 4 formants and tracks of artificially set harmonics with the fundamental frequency of 125 Hz.

Техника компрессии речи через сжатие изображений ее спектрограмм

Сжатие-восстановление речи через сжатие-восстановление изображений ее спектрограмм позволяет получить синтезированную речь более высокого качества звучания, чем только по параметрам главных опорных синусоид. Здесь может применяться большое количество алгоритмов сжатия изображений (конвертации форматов) как с частичной потерей информации, так и без.

В результате экспериментов по преобразованию форматов без потери информации получено, что наибольший эффект сжатия исходного полутонового изображения спектрограммы достигается при его конвертации в формат *.flif. Этот же метод показал достаточную эффективность и при конвертации бинарных изображений спектрограмм.

В табл. 1 приведено сравнение коэффициентов и степени сжатия известных форматов при их применении без потерь к бинарному изображению динамической спектрограммы *5-ти* речевых сигналов длительностью *30* секунд каждый.

Коэффициент сжатия рассчитывался по формуле:

$$R = \frac{L_{\text{исх}}}{L_{\text{сж}}}, \tag{3}$$

где $L_{\text{исх}}$ – объем данных исходного изображения, а $L_{\text{сж}}$ – объем сжатых данных.

Степень сжатия *r*, характеризующая относительное уменьшение объема данных:

$$r = \frac{L_{\text{исх}} - L_{\text{сж}}}{L_{\text{исх}}} \times 100\%. \tag{4}$$

Для черно-белого изображения более лучшие результаты удалось достичь, используя его преобразование в формат *.jbg. Данный формат хорошо зарекомендовал себя для сжатия факсимильных изображений, но на практике показал эффективность и при сжатии других типов бинарных изображений, в частности бинарных спектрограмм. Сопоставимый результат показал и формат *.flif. Однако преобразование *.jbg по сравнению с *.flif является менее ресурсоемким по времени.

Также близкий результат к формату *.jbg показало преобразование в формат *.tif с параметрами сжатия CCITT Group 4 (G 4), который тоже разрабатывался и использовался для сжатия факсимильных черно-белых изображений.

Таблица 1. Сравнение алгоритмов конвертации бинарного изображения спектрограммы

Формат исходных изображений	$L_{\text{исх}}$, Кбайт	Форматы конвертации	$L_{\text{сж}}$, Кбайт	R	r, %
*.bmp (1 bpp)	287	flif	18	15,94	93,73%
		gif	49	5,86	82,93%
		jbg	17	16,88	94,08%
		jif	50	5,74	82,58%
		png	48	5,98	83,28%
		tga	286	1,00	0,35%
		tif (G 4)	24	11,96	91,64%
		webp	35	8,20	87,80%

Предлагаемая техника позволяет эффективно управлять пропускной способностью канала передачи-хранения в диапазоне *4,8–16 Кбод*. В зависимости от ее требуемого значения в канал передаются параметры сжатия полутонового изображения спектрограммы РС, а в случае необходимости серьезного понижения пропускной способности – бинарного изображения спектрограммы.

Качественный синтез речи по изображению восстановленной спектрограммы в рамках данной техники возможен посредством использования алгоритма Гриффина-Лима. Однако он весьма сложен и требует большого количества итераций, несмотря на появление его ускоренных модификаций [9, 13–15]. В тоже время благодаря неитеративному характеру алгоритм Маккалая-Кватьери работает очень быстро и подходит для протяжённых аудиосигналов.

Очевидным недостатком алгоритма Гриффина-Лима является заполнение фазовой составляющей синтезируемого РС случайными значениями на первой итерации, в результате у нового сигнала имеются недостатки в виде артефактов и некоторой неестественности звучания голоса («железности»).

Безитерационная однопроходная модель Маккалая-Кватьери [3], проводя синтез по параметрам пиков опорных синусоид, позволяет достаточно хорошо формировать фазовый спектр РС, однако за счет грубости представления амплитуд и частот пиков восстановленный (синтезированный) по Маккалая-Кватьери РС, несмотря на имеющуюся схожесть звучания с исходным, несколько сильнее отличается от оригинала чем речь, синтезируемая по алгоритму Гриффина-Лима.

Поэтому был предложен вариант смешанной техники. После передачи по каналу на восстановленной спектрограмме выделяются параметры пиков, по которым проводится первая итерация процедуры синтеза нового РС в гребенке генераторов синусоид по алгоритму Маккалая – Квартьери [2, 3] (рис. 4) и с последующими итерациями и синтезом РС уже по алгоритму Гриффина-Лима [13–15].

Для оценки качества так восстановленного РС использовался метод PESQ (Perceptual Evaluation of Speech Quality – ITU-T recommendation P.862 (02/01)), входящий в семейство стандартов, включающих методику тестирования для автоматизированной оценки качества речи. Результаты оценки с использованием PESQ отражены в табл. 2.

Восстановленные сигналы, полученные по смешанной технике спектральной инверсии на скоростях передачи *4,8–16 КБод* обладали наилучшим качеством звучания.

Таким образом, результаты итеративных алгоритмов инверсии спектрограмм с восстановлением фазы, включая алгоритм Гриффина-Лима, могут быть существенно улучшены если инициализировать их оценкой фазы, полученной изначально безитерационным алгоритмом.

Таблица 2. Оценка качества восстановленной речи методом PESQ

№ п/п	Наименования алгоритма синтеза речеподобного сигнала	Количество итераций алгоритма Гриффина-Лима	Оценка PESQMOS
1	Алгоритм Маккалая-Кватьери	безитерационный	3,219
2	Алгоритм Гриффина-Лима	~150	3,226
3	Алгоритм Гриффина-Лима с инициацией фазы синтеза РС по алгоритму Маккалая-Кватьери	~20	3,231

Заключение

В данном исследовании предлагается метод аудио кодирования с адаптацией к изменениям пропускной способности речевого канала связи, позволяющий плавно регулировать скорость передачи-хранения РС в широком диапазоне (*1,2–16 Кбит/с*) и с сохранением сходства звучания с оригиналом, попутно избавляясь от присутствующих шумов и помех.

Предлагаемый аудиокодек основан на синусоидальной Гауссовой модели анализа/синтеза речи, где суперпозиция взвешенных колоколом Гаусса гармонических компонент, применима для всех видов речевых фреймов, и универсальных методах построения и обработки изображений узкополосных динамических спектрограмм.

Сжатие-восстановление речи через прямое сжатие-восстановление изображений ее полутоновых спектрограмм хорошо показало себя для скоростей передачи свыше *4,8 Кбод.* Качество восстановленной речи практически не отличается от исходной.

Для низкоскоростного кодирования (*4,8–2,4 Кбод*) стоит применять специальные методы выделения и формирования пиковых параметров синусоидальной Гауссовой модели.

На скоростях ниже *2,4 Кбод* предпочтительней кодировать параметры глобальных максимумов спектра, определяемых движением формант, и использовать искусственно восстанавливаемую фазу, что не потребует дополнительных бит при параметрическом описании участков РС.

Предложенные методы сжатия-восстановления речи через обработку изображений спектрограмм могут одновременно выполняться с удалением параметров шумов и помех и/или с заменой испорченных шумом участков спектрограмм на эталонные из базы данных голоса диктора для обеспечения высокого качества восстановленной речи.

В качестве преимуществ исследованных и разработанных здесь подходов к кодированию речи следует отметить, что в предложенном аудиокодеке:

- задействованы эффективные в вычислительном плане процедуры и алгоритмы;
- речевой сигнал реконструируется и восстанавливается с высоким качеством;
- улучшается качество речевого сигнала, испорченного аддитивным шумом;
- используемые методы и модели помехоустойчивы, не зависят от высоты тона (основной частоты);
- могут одновременно выполняться процедуры обнаружения и исправления ошибок.

Как направление дальнейших исследований следует оценить возможность создания сверх низкоскоростного кодека путем включения в состав кодера блока распознавания речи с кодировкой распознанного текста, его передачей и восстановлением по нему на приемном конце канала речевой связи спектрограммы и синтеза по ней речи по известной технологии «text-to-speech» голосом клона конкретного диктора, используя заранее сформированную голосовую базу данных.

Основной задачей здесь будет разработка и тестирование быстрого и эффективного алгоритма спектральной инверсии без задержки синтеза для работы в устройствах реального времени. Как вариант ее решения: использование надёжных и эффективных итеративных алгоритмов восстановления фазы, которые могут быть инициализированы фазой, изначально полученной безитерационным алгоритмом инверсии спектрограмм.

СПИСОК ЛИТЕРАТУРЫ:

1. Spanias, Speech Coding: A Tutorial Review, Proc. of the IEEE, Vol. 82. No. 10. P. 1541–1582, Oct. 94. DOI: http://dx.doi.org/10.1109/5.326413.
2. Samer J. Alabed, Eyad A. Ibrahim. A new sinusoidal speech coding technique with speech Enhancer at low bit rates. International Journal of Electronics and Communication Engineering & Technology (IJECET). Vol. 5, Issue 4, April (2014). P. 07–18. URL:https://www.academia.edu/17677709/2_a_new_sinusoidal_speech_coding_technique_with_speech_enhancer_at_low_bit_rates (дата обращения: 08.11.2021).
3. R.J. McAulay and T.F. Quatieri, Speech Analysis/Synthesis Based on a Sinusoidal Representation, IEEE Trans. On ASSP. Vol. ASSP-34. No. 4. P. 744–754, August 1986. DOI: http://dx.doi.org/10.1109/TASSP.1986.1164910 (дата обращения: 08.11.2021).
4. Griffin D.W. and Lim J.S., Signal estimation from modified short-time Fourier transform, IEEE Transactions on Acoustics, Speech and Signal Processing. 1984. P. 236–243. DOI: http://dx.doi.org/10.1109/TASSP.1984.1164317.
5. Дворянкин Сергей В.; Дворянкин Никита С.; Устинов Роман А. Развитие технологий образного анализа-синтеза акустической (речевой) информации в системах управления, безопасности и связи. Безопасность информационных технологий, [S.l.]. Т. 26, № 1. С. 64–76, 2019. ISSN 2074-7136. DOI: http://dx.doi.org/10.26583/bit.2019.1.07.
6. Дворянкин Сергей В. и др. Системное моделирование речеподобных сигналов и его применение в сфере безопасности, связи и управления. Безопасность информационных технологий, [S.l.]. Т. 26, № 4. С. 101–119, 2019. ISSN 2074-7136. DOI: http://dx.doi.org/10.26583/bit.2019.4.08.
7. Dan Ellis. Robust Landmark-Based Audio Fingerprinting. MATLAB Central File Exchange. Retrieved November 2, 2021. URL: https://www.mathworks.com/matlabcentral/fileexchange/23332-robust-landmark-based-audio-fingerprinting (дата обращения: 08.11.2021).
8. Дворянкин С.В., Михайлов Д.М., Панфилов Л.А., Бонч-Бруевич А.М., Козлачков С.Б., Насенков И.Г. Интерпретация и контурный анализ спектрограмм звуковых сигналов в процессе их шумоочистки. Проблемы информационной безопасности. Компьютерные системы. 2015. № 3. С. 88–99. URL: https://jisp.ru/volume/2015/ (дата обращения: 25.11.2021).
9. Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin–lim iteration. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). P. 61–65, 2019. DOI: http://dx.doi.org/10.1109/ICASSP.2019.8682744.
10. Женило В. Р. Компьютерная фоноскопия. М.: Изд-во Академии МВД России, 1995. – 207 с.
11. Алюшин В.М., Дворянкин С.В. Технологии образного анализа в задачах цифровой обработки речевой информации. Научная визуализация. 2013. Т. 5. № 3. С. 75–88. URL: http://sv-journal.org/2013-3/06/index.html (дата обращения: 08.11.2021).
12. Дворянкин С.В., Устинов Р.А. Методы синтеза речеподобных сигналов по изображениям динамических спектрограмм / II Межведомственная научно-практическая конференция «Телекоммуникации и кибербезопасность: специальные системы и технологии». Сб. тр. // Под ред. засл. деятеля науки РФ, почетного радиста РФ, дтн, проф. В.А. Цимбала и дтн, проф. О.И. Атакищева. Серпухов: МОУ «ИИФ», 2020. Т. 3. С. 170–180.
13. Yoshiki Masuyama, Kohei Yatabe, and Yasuhiro Oikawa. Griffin–lim like phase recovery via alternating direction method of multipliers. IEEE Signal Processing Letters, 26(1):184–188, 2018. DOI: http://dx.doi.org/10.1109/LSP.2018.2884026.
14. S.Ö. Arık, H. Jun and G. Diamos. Fast Spectrogram Inversion Using Multi-Head Convolutional Neural Networks, in IEEE Signal Processing Letters. Vol. 26, no. 1. P. 94–98, Jan. 2019. DOI: http://dx.doi.org/10.1109/LSP.2018.2880284.
15. Sharma, A., Kumar, P., Maddukuri, V. et al. Fast Griffin Lim based waveform generation strategy for text-to-speech synthesis. Multimed Tools Appl 79, 30205–30233 (2020). DOI: http://dx.doi.org/10.1007/s11042-020-09321-7.

REFERENCES:

[1] Spanias, Speech Coding: A Tutorial Review, Proc. of the IEEE, Vol. 82. No. 10. P. 1541–1582, Oct. 94. DOI: http://dx.doi.org/10.1109/5.326413.

[2] Samer J. Alabed, Eyad A. Ibrahim. A new sinusoidal speech coding technique with speech Enhancer at low bit rates. International Journal of Electronics and Communication Engineering & Technology (IJECET). Vol. 5, Issue 4, April (2014). P. 07–18. URL:https://www.academia.edu/17677709/2_a_new_sinusoidal_speech_coding_technique_with_speech_enhancer_at_low_bit_rates (accessed: 08.11.2021).

[3] R.J. McAulay and T.F. Quatieri, Speech Analysis/Synthesis Based on a Sinusoidal Representation, IEEE Trans. On ASSP. Vol. ASSP-34. No. 4. P. 744–754, August 1986. DOI: http://dx.doi.org/10.1109/TASSP.1986.1164910.

[4] Griffin D.W. and Lim J.S., Signal estimation from modified short-time Fourier transform, IEEE Transactions on Acoustics, Speech and Signal Processing. 1984. P. 236–243. DOI: http://dx.doi.org/10.1109/TASSP.1984.1164317.

[5] Dvoryankin Sergey V.; Dvoryankin Nikita S.; Ustinov Roman A. Improvement of image analysis/synthesis technologies of acoustic (speech) information for the control, safety and communication systems. IT Security (Russia), [S.l.]. Vol. 26, no. 1. P. 64–76, 2019. ISSN 2074-7136. DOI: http://dx.doi.org/10.26583/bit.2019.1.07 (in Russian).

[6] Dvoryankin Sergey V. et al. Speech-like signal system modeling and its application in the field of security, communication and control access. IT Security (Russia), [S.l.]. Vol. 26, no. 4. P. 101–119, 2019. ISSN 2074-7136. DOI: http://dx.doi.org/10.26583/bit.2019.4.08 (in Russian).

[7] Dan Ellis. Robust Landmark-Based Audio Fingerprinting. MATLAB Central File Exchange. Retrieved November 2, 2021. URL: https://www.mathworks.com/matlabcentral/fileexchange/23332-robust-landmark-based-audio-fingerprinting (accessed: 08.11.2021).

[8] Dvoryankin S.V., Mikhailov D.M., Panfilov L.A., Bonch-Bruevich A.M., Kozlachkov S.B., Nasenkov I.G. Interpretation and contour analysis of spectrograms of audio signals in the process of their noise cleaning. Problems of information security. Computer systems. 2015. № 3. С. 88–99. URL: https://jisp.ru/volume/2015/ (accessed: 25.11.2021) (in Russian).

[9] Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin–lim iteration. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). P. 61–65, 2019. DOI: http://dx.doi.org/10.1109/ICASSP.2019.8682744.

[10] Zhenilo V.R. Computer phonoscopy. M.: Publishing house of the Academy of the Ministry of Internal Affairs of Russia, 1995. – 207 p. (in Russian).

[11] Alyushin V.M., Dvoryankin S.V. Technologies of figurative analysis in the tasks of digital processing of speech information. Scientific visualization. 2013. T. 5. № 3. С. 75–88. URL: http://sv-journal.org/2013-3/06/index.html (accessed: 08.11.2021) (in Russian).

[12] Dvoryankin S.V., Ustinov R.A. Methods for the synthesis of speech-like signals from images of dynamic spectrograms. II Interdepartmental Scientific and Practical Conference "Telecommunications and Cybersecurity: Special Systems and Technologies". Collection of works. Edited by Honored Scientist of the Russian Federation, Honorary Radio Operator of the Russian Federation, Doctor of Technical Sciences, prof. V.A. Tsimbala and Doctor of Technical Sciences, prof. O.I. Attackischeva. Serpukhov: MOU "IIF", 2020. Vol. 3. P. 170–180 (in Russian).

[13] Yoshiki Masuyama, Kohei Yatabe, and Yasuhiro Oikawa. Griffin–lim like phase recovery via alternating direction method of multipliers. IEEE Signal Processing Letters, 26(1):184–188, 2018. DOI: http://dx.doi.org/10.1109/LSP.2018.2884026.

[14] S.Ö. Arık, H. Jun and G. Diamos. Fast Spectrogram Inversion Using Multi-Head Convolutional Neural Networks, in IEEE Signal Processing Letters. Vol. 26, no. 1. P. 94–98, Jan. 2019. DOI: http://dx.doi.org/10.1109/LSP.2018.2880284.

[15] Sharma, A., Kumar, P., Maddukuri, V. et al. Fast Griffin Lim based waveform generation strategy for text-to-speech synthesis. Multimed Tools Appl 79, 30205–30233 (2020). DOI: http://dx.doi.org/10.1007/s11042-020-09321-7.

Поступила в редакцию – 01 ноября 2021 г. Окончательный вариант – 25 ноября 2021 г.
Received – November 01, 2021. The final version – November 25, 2021.