БЫСТРЫЙ СИНТЕЗ АУДИОСИГНАЛОВ ПО ИЗОБРАЖЕНИЯМ СПЕКТРОГРАММ В ЗАДАЧАХ ЗАЩИТЫ РЕЧЕВОЙ ИНФОРМАЦИИ

Дворянкин С. В.[1], Дворянкин Н. С.[2], Алюшин А. М.[3]

DOI: 10.21681/2311-3456-2024-5-34-46

Цель исследования: *разработка методов и алгоритмов инверсии спектрограмм: синтеза волновой формы сигнала по заранее известным данным его амплитудных спектральных разверток в отсутствии информации о фазе, – для генерации в реальном масштабе времени аудиосигналов с заданными частотно-временными свойствами с их последующем применением в системах защиты речевой информации.*

Методы исследования: *прикладного системного анализа, цифрового спектрально-временного анализа, цифровой обработки сигналов и изображений, образного анализа сонограмм.*

Результаты исследования: *предложены методы и алгоритмы синтеза звуковых и речевых сигналов по априори заданной спектрограмме, реализуемые в рамках концепции образного анализа-синтеза, работающие в реальном масштабе времени и обеспечивающие хорошие качественные оценки фазы пиковых значений спектральных срезов за один полностью детерминированный проход. Могут использоваться самостоятельно или для получения начальных оценок фазы для улучшения результатов итеративных алгоритмов типа Гриффина-Лима и др. Получаемые по обработанным изображениям спектрограмм оценки позиций и фазы спектральных пиков определяются точнее с помощью квадратичной интерполяции, а расчет приращения фазы по шагам времени ведется в специально введенном фазовом аккумуляторе, не требуя вычисления арктангенсов.*

Научная новизна: *предложен новый метод инверсии спектрограмм на основе рассечения-разнесения образа исходной спектрограммы для получения более точных спектральных описаний, синтезированного по ней аудиосигнала, лучше соответствующих оригиналу, чем у известных итерационных методов спектральной инверсии.*

Практическая значимость: *разработан эффективный, с точки зрения вычислений, алгоритм реального времени для однопроходной инверсии спектрограмм. Полученные результаты позволят расширить возможности существующих систем защиты речевой информации и проектировать более эффективные на основе изложенных подходов.*

Ключевые слова: *информационная безопасность, инверсия спектрограмм, образный анализ, защита от несанкционированного доступа, речеподобный сигнал, синусоидальная модель речи.*

Введение

Анализ существующих методов и средств обработки и защиты речевой информации (ЗРИ) от НСД показывает, что все они, так или иначе связаны с трансформацией, модификацией и-или заменой спектральных характеристик исходного речевого сигнала (РС), прежде всего с изменениями динамических разверток амплитудного спектра – спектрограмм[4] [1–8].

Сегодня амплитудные спектрограммы (для речи сонограммы) широко используются для представления, визуализации и выполнения операций над сигналами в частотной области. Приложения с их участием включают, но не ограничиваются следующими областями: перевод текста в речь, техническое маскирование (закрытие) речи, распознавание речи, генерация активных речеподобных помех для выделенных помещений, улучшение качества звука, акустическая стеганография, аудиокодеки и сжатие речи, изменение масштаба звучания по времени, изменение высоты тона, конвергенция и клонирование голоса, идентификация диктора, шумоподавление, реконструкция искаженных фонограмм и др [5,6,7]. [1–8].

Во многих приложениях, в том числе для ЗРИ, необходимы анализ и модификация изображений

1 Дворянкин Сергей Владимирович, доктор технических наук, профессор, профессор кафедры стратегических информационных исследований НИЯУ МИФИ, заведующий лабораторией защиты и обработки аудиовизуальной информации МГЛУ, г. Москва, Россия. E-mail: svdvoryankin@mephi.ru, https://orcid.org/0000-000-6908-0676

2 Дворянкин Никита Сергеевич, аспирант НИЯУ МИФИ, г. Москва, Россия. E-mail: nik.dvrn@gmail.com

3 Алюшин Александр Михайлович, старший преподаватель кафедры информатики и процессов управления НИЯУ МИФИ, научный сотрудник лаборатории защиты и обработки аудиовизуальной информации МГЛУ, г. Москва, Россия. E-mail:alyshin@list.ru

4 Барсуков В. С., Дворянкин С. В., Шеремет И. А. Безопасность связи в каналах телекоммуникаций / М.: НИФ «Электронные знания», 1992. 122 с.

5 Дворянкин С. В. Цифровая шумоочистка аудиоинформации. Под ред. А. В. Петракова. М.: ИП РадиоСофт, 2011. 208 с.

6 Дворянкин С. В., Макаров Ю. К., Хорев А. А. Обоснование критериев эффективности защиты речевой информации от утечки по техническим каналам // Защита информации. Инсайд. – № 2 (14). Март-апрель 2007. С. 18–25.

7 Петраков А. В., Лагутин В. С. Утечка и защита информации в телефонных каналах / М.: Энергоатомиздат, 1998. 317 с.

спектрограмм, полученных в результате кратковременного преобразования Фурье (КПФ) аудиосигналов, с последующим переходом к новому сигналу во временной области с так заданными свойствами. Совокупность указанных модулей и процедур, а именно: построение спектрограмм посредством КПФ, их обработка, трансформация или замена, в том числе методами искусственного интеллекта (ИИ), а также инверсия спектрограмм для получения нового речеподобного сигнала (РПС) с нужными характеристиками – составляют основу технологии «звук – изображение – звук» реализуемой в системе образного анализа – синтеза (ОАС), представленной на рис. 1, активно и успешно продвигаемой в решении ряда задач ЗРИ [1–8].

Рис. 1. Общая схема системы образного анализа-синтеза речи для технологии «звук-изображение-звук» при моделировании методов и средств ЗРИ в РМВ

Крайний модуль системы ОАС: «инверсия спектрограмм» или вокодер, – вычислительно ресурсоемкий, требует значительного времени на свою реализацию. Выходной сигнал РПС как правило имеет недостатки и по качеству звучания. В современных исследованиях функции вокодера перекладываются на нейросетевые решения и алгоритмы, что значительно суживает области применения ОАС в автономных системах ЗРИ, особенно работающих в режимах масштаба времени близких к реальному (РМВ).

В связи с этим разработка и исследование эффективного для РМВ алгоритма однопроходной инверсии спектрограмм в рамках концепции ОАС для решения задач защиты речевой информации остается весьма актуальной.

Образный анализ-синтез акустического речевого сигнала

Рассмотрим основные модули ОАС подробнее (рис.1).

Модуль построения спектрограмм

Спектрограмма сигнала представляет собой последовательность спектральных срезов, получаемых в ходе выполнения скачущего или скользящего кратковременного преобразования Фурье (КПФ). Уравнение КПФ известно как:

$$X(mS,\varpi) = \sum_{n=-\infty}^{\infty} x(n)w(n - mS)e^{-j\varpi n} \qquad (1)$$

Мгновенный амплитудный спектр как результат КПФ может быть представлен в виде модуля $|\ X(mS,\varpi)\ |$ этого преобразования, где w – окно анализа, S – шаг анализа (скачка) по оси времени, ϖ – круговая частота и m – индекс текущего кадра (фрейма) исходного сигнала, над которым делается КПФ.

На практике процедура вычисления КПФ состоит в том, чтобы разделить сигнал длительного времени на более короткие, перекрывающиеся сегменты равной длины, а затем вычислить через быстрое преобразование Фурье БПФ) амплитудный и фазовый спектры, на каждом выделенном коротком сегменте (фрейме).

Ансамбль последовательно получаемых неотрицательных амплитудных спектральных срезов позволяет рассматривать спектрограмму как некое изображение (графический образ), где в уровнях одного выбранного цвета (например, серого) на частотно-временной сетке отображаются мощностные характеристики звукового сигнала и «следы» (треки) элементарных гармоник его составляющих (рис. 2) [4, 6].

Модуль обработки изображений спектрограмм

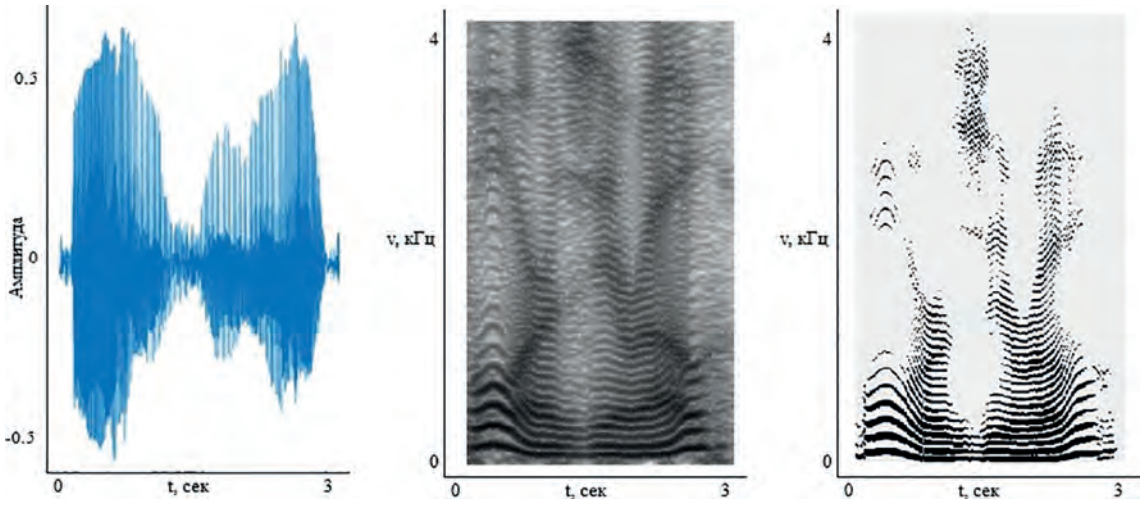
Современная обработка спектрограмм для различных приложений предполагает применение не только методов цифровой обработки изображений (ЦОИ) и сигналов, но и методов машинного обучения, распознавания образов, эффективных решений искусственного интеллекта (ИИ).

Отметим, что без потери информативности при переходе от временной к частотной области анализа возможны два типа представления спектрограммы: полутоновая и бинарная. Соответственно разные методы обработки изображений могут быть применены для решения той или иной задачи. Так хорошо проработанные давно известные методы обработки бинарных изображений показывают неплохие результаты в решении задач сжатия речи через сжатие её бинарных сонограмм [4, 6].

А реконструкция спектрограмм искаженных РС с помощью нейросетей и речевой базы данных целевого диктора позволяет решать задачи обработки и защиты речевой информации, ранее трудно решаемые или нерешаемые совсем. Например, речевая подпись, речевая реабилитация, адаптивная речеподобная помеха, стойкая к шумоочистке, и др.

Модуль инверсии спектрограммы (вокодер)

Как правило, амплитудные спектрограммы обрабатываются отдельно от фазовых составляющих частотных компонентов. Поэтому в некоторых приложениях, в частности в области ЗРИ, часто необходима инверсия спектрограммы, как процесс реконструкции волновой формы сигнала во временной области по его уже имеющейся заданной спектрограмме,



А) Б) В)

Рис. 2. Осциллограмма и спектрограммы РС: а) – волновая форма РС; б) – узкополосная полутоновая спектрограмма (гармоническая и формантная структура вместе); в) – бинарная спектрограмма с треками локальных максимумов узкополосных составляющих речи.

для которого требуется оценка фаз частотных составляющих спектра.

Если даже никаких изменений в исходных изображениях спектрограмм не проводилось, то волновая форма созданных по ним РПС может не совпадать с оригинальной и даже существенно от неё отличаться из-за различных начальных фаз в моделях РС, используемых в процедуре синтеза. А звучание оригинального и синтезированного по спектрограмме сигналов тем не менее будет идентичным, поскольку слух почти не восприимчив к фазам базовых гармоник.

В процессе аудио обработки, фазы либо теряются, либо становятся некорректными, либо их просто не существует для искусственно созданных спектрограмм. Таким образом, задача модуля состоит в том, чтобы использовать полученные посредством КПФ и модифицированные с помощью ИИ и ЦОИ амплитудные спектральные описания для генерации сигнала, спектр которого наилучшим образом будет соответствовать оригинальным спектрограммам.

Рассмотрим подробнее вопросы создания модуля однопроходной инверсии спектрограмм (вокодера) реального времени для решения задач ЗРИ. В качестве прототипов, подлежащих уточнению и совершенствованию, возьмем за основу и сориентируемся на алгоритмы инверсии спектрограмм и синтеза РПС с заданными спектральными характеристиками, приведенные в работах[8,9,10] [4, 6, 7].

Анализ существующих алгоритмов инверсии спектрограмм

В блоке вокодера (рис. 1) в основе преобразования (инверсии) изображения сгенерированной спектрограммы в новый звуковой или речевой сигнал от целевого диктора довольно часто используется итерационный алгоритм Гриффина-Лима (GLA) или ему подобные или производные [9–25], имеющие схожие недостатки по скорости и качеству синтеза нового РПС. Рассмотрим алгоритм Гриффина-Лима (GLA)[11].

Начиная с первоначальной оценки $X^0(n)$ исходного сигнала во временной области $x(n)$, каждая итерация алгоритма GLA итеративно обновляет оценку:

$$x^{i+1}(\mathrm{n}) = \frac{\sum_{m=-\infty}^{\infty} w(n - mS)\,\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} \hat{X}^i(mS,\omega)\, e^{-jwn}\, d\omega}{\sum_{m=-\infty}^{\infty} w^2(n - mS)}. \quad (2)$$

Здесь $\hat{X}^i(mS,\omega)$ есть результат КПФ от $x^i(n)$ со следующим ограничением на величину:

$$\hat{X}^i(mS,\omega) = X^i(mS,\omega)\frac{|\,X(mS,\omega)\,|}{|\,X^i(mS,\omega)\,|}, \quad (3)$$

где $|\,X(mS,\omega)\,|$ – модуль спектра в результате КПФ от исходного сигнала $x(n)$, $|\,X^i(mS,\omega)\,|$ – модуль КПФ i-й оценки $x^i(n)$, а S соответствует размеру шага анализа (сдвига) окна.

То есть, каждая новая итерация дает новый набор фаз, которые сочетаются с исходным спектром амплитуд для проведения следующей итерации.

Расстояние или мера близости часто рассчитывается как квадратичная ошибка между исходной

8 Beauregard, Gerald Harish, Mithila Wyse, Lonce Single Pass Spectrogram Inversion 2015/07/01, pp. 427–431. DOI – 10.1109/ICDSP.2015.7251907

9 R. Decorsiere, P. L. Sondergaard, E. N. MacDonald, and T. Dau, «Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations» Audio Speech Lang. Process. IEEEACM Trans. On, vol. 23, no. 1, pp. 46–56, 2015.

10 Дворянкин С. В. Речевая подпись. М.: РИО-МТУСИ. 2003. – 184 с.

11 D. Griffin and J. Lim, «Signal estimation from modified short-time fourier transform» Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 32, no. 2, pp. 236–243, 1984.

и реконструированной сигнальными спектрограммами с использованием показателя SER (Signal Error Ratio). Одним из достоинств алгоритма Гриффина-Лима[12] [15, 16] заключается в том, что он монотонно увеличивает SER с каждой новой итерацией. Для получения качественного синтезируемого сигнала на практике таких итераций должно быть около сотни, что требует существенного вычислительного ресурса.

Существенным недостатком является то, что оценка фазы для текущего кадра зависит от всех будущих и всех прошлых кадров исходного сигнала, а звучание нового синтезированного сигнала имеет металлизированные оттенки [8, 9] из-за размытости спектра на верхних частотах. Таким образом, этот метод уже по своей сути, изначально не является методом реального времени, что крайне востребовано в автономных системах ЗРИ.

Кроме того, отмечается, что для GLA важно выбрать подходящие начальные оценки фазовых компонент поскольку различающиеся начальные оценки дают разные результаты, и нет гарантии, что будет достигнуто оптимальное решение [13, 14, 15].

Нейросетевые алгоритмы восстановления фазы также ресурсоемкие и качество звучания синтезированного сигнала часто оставляет желать лучшего.

Поэтому для приложений РМВ в области ЗРИ предлагается разработать алгоритм однопроходной инверсии спектрограммы (ОИС), лишенный указанных недостатков алгоритма «GLA» и ему подобных, с использованием найденных зависимостей между амплитудным и фазовым спектрами РС[13], позволяющих восстанавливать фазовый спектр по амплитудному с точностью до начальной фазы и амплитудный спектр по фазовому с точностью до постоянного множителя.

Для этого сначала уточним используемую модель РС пофреймно рассматриваемого в виде суперпозиции узкополосных опорных синусоидальных сигналов[14] (синусоидальная модель [4–6]).

Синусоидальная модель РС и методы синтеза по локальным максимумам спектральных срезов

Во многих приложениях ЗРИ используются Гильбертовские модели, где РС рассматривается как произведение неотрицательной огибающей на косинус фазы. Согласно уточненной синусоидальной модели исходный РС при длительности анализируемого фрейма речи менее *40 мс* (оптимально 6–*8 мс*) [4–6] может быть представлен как:

$$x(n) = \sum_{k=1}^{K} A_k \, e^{-n^2/\sigma_{nk}} \cos(\omega_k n + \theta(n) + \varphi_{0k}), \qquad (4)$$

где n – номер временного отсчета; K – количество значимых синусоид для текущего фрейма; A_k – амплитуда k-й синусоиды; ω_k –круговая частота и φ_{0k} – начальная фаза k-й синусоиды, σ – эффективная ширина окна функции Гаусса, $\theta(n)$ – нелинейная часть фазы.

Достоверность данного описания может быть проверена путем сравнения фазовых значений элементарной гармоники из соотношения (4) на двух соседних спектральных срезах по фазограмме[15]:

$$\lim_{\Delta t \to 0} (\varphi_{i+1,k} - \varphi_{i,j} - \omega_i \Delta t) = 0 \qquad (5)$$

где $\varphi_{i,j}$ – значение фазы гармоники с частотой, соответствующей j-ому элементу ДПФ на i-ом спектральном срезе, $\varphi_{i+1,k}$ – значение фазы этой же гармоники, соответствующей k-ому элементу ДПФ на $(i+1)$-ом спектральном срезе, Δt – временной интервал между двумя соседними спектральными срезами.

Данное описание РС по формуле (4) обладает рядом преимуществ:

1. Здесь, с одной стороны, привнесённое окно Гаусса применяется для сглаживания краевых эффектов на границах фреймов, а с другой – позволяет рассматривать исходный речевой сигнал на каждом фрейме как суперпозицию узкополосных сигналов или вейвлетов Морле (синусоид взвешенных окном Гаусса).
2. Для синтеза речи по изображению спектрограммы достаточно учитывать только локальные максимумы спектрального среза, полученного в результате ДПФ.

   Действительно, в соответствии со следствиями преобразования Фурье перемножение функции окна на гармонический сигнал во временной области приводит к сдвигу образа окна на частоту этой синусоиды в частотной области. А образ окна Гаусса также является окном Гаусса. Следовательно локальные максимумы (ЛМ) или пиковые бины на столбцах (срезах) изображений спектрограмм с параметрами $\{A_i; \omega_i; \varphi_{0i}\}$ полностью определяют базовые узкополосные составляющие (4) анализируемого фрейма, по которым они могут быть восстановлены в составе нового синтезируемого по ним РС.
3. Появляется возможность описывать РС в виде (4) как на вокализованных, так и на невокализованных участках как суперпозицию элементарных базовых гармоник.
4. Для восстановления звукового сообщения только по изображению спектрограммы можно синтетически рассчитать фазу сигнала для ЛМ спектральных срезов, что может быть использовано при

12 D. Griffin and J. Lim, «Signal estimation from modified short-time fourier transform» Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 32, no. 2, pp. 236–243, 1984.

13 Дворянкин С. В. Речевая подпись. М.: РИО-МТУСИ. 2003. – 184 с.

14 R. McAulay and T. Quatieri, Speech analysis/Synthesis based on a sinusoidal representation, in IEEE Transactions on Acoustics, Speech, and Signal Processing. V. 34, no. 4. P. 744–754, August 1986.

15 Дворянкин С. В. Речевая подпись. М.: РИО-МТУСИ. 2003. – 184 с.

синтезе сигнала без использования оригинальных фазовых значений.

Исходя из этих преимуществ можно определить выражение (4) в качестве базового описания РС и предусмотреть возможность следующих видов синтеза РПС по ЛМ, найденным на исходной спектрограмме с оригинальной и искусственной фазой:

- ✓ синтез на основе обратного БПФ по трекам ЛМ синусоидальных составляющих на срезах, полученных с использованием КПФ;
- ✓ синтез по уточненным позициям ЛМ на спектральном срезе с использованием генераторов синусоидальных колебании;
- ✓ синтез по ЛМ базисных функций преобразования Фурье с рассечением-разнесением (прореживанием) спектрограммы.

Указанные методы синтеза будут подробно рассмотрены ниже.

Протяжка фазы для спектральных локальных максимумов

При отсутствии данных о фазе сигнала, например, для шумоочистки изображения его сонограммы и последующего синтеза по ней разборчивого РС, необходимо провести «протяжку» фазы, суть которой заключается в следующем:

а) в начальный момент времени полагаем фазу каждой гармоники равной нулю или случайному значению (на разборчивость речи это не влияет);

б) в каждый текущий момент времени фаза гармоники с номером i находится по формуле

$$\varphi_i(t + \Delta t) = \varphi_j(t) + \Delta t \frac{2\pi}{N} i, \text{ где } \arg\max_{j \in \{i-3,\ i+3\}} | X_j(t) | \qquad (6)$$

в) если на предыдущем временном срезе пределах створа наблюдения не найден ЛМ гармоники текущего среза, то фаза берется нулевой/случайной.

Таким образом становится возможным провести синтез речи только по изображению сонограммы, не имея в наличии оригинальной фазограммы сигнала.

Изменения в спектральных компонентах треков ЛМ узкополосных (синусоидальных) составляющих отслеживаются на обрабатываемой спектрограмме с использованием понятий «рождение», «жизнь» и «смерть», лежащих в основе принятого синусоидального представления (4) для РС[16].

Иллюстрация результатов процедур спектральных изменений в виде процессов «рождения», «смерти» и «жизни» (продолжения треков ЛМ) с учетом возможных положений пиковых бинов на последовательности спектральных срезов показана на рис. 3.

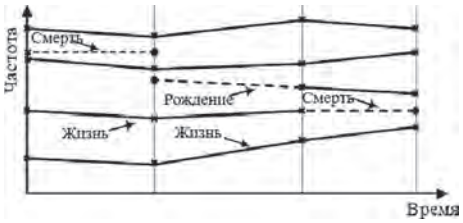


Рис. 3. Сопоставление на спектральных срезах треков локальных максимумов элементарных опорных гармоник и процедур «рождения-жизни-смерти»

На этапе синтеза, для каждого трека (следа) или контура базовой синусоиды, определенного на изображении спектрограммы (например, траектории линий гармоник основного тона на вокализованных фреймах), к заданным параметрам частоты и амплитуды трека будет присоединяться выбранная фазовая функция (оригинальная или искусственная в зависимости от задачи), необходимая для разворачивания и интерполяции фазы, построенная таким образом, чтобы фазовый след был максимально гладким [4–6].

Детализированная система образного анализа-синтеза речевого сигнала на основе синусоидальной узкополосной модели с учетом свойств слуха в составе модулей КПФ, обработки локальных максимумов треков базовых синусоид и инверсии спектрограмм показана на рис. 4.

В зависимости от решаемой задачи для каждого ЛМ на текущем спектральном срезе выбирается фазовая функция, либо как исходная, вычисленная по комплексному спектру исходного сигнала в ходе КПФ, либо искусственная, вычисленная на основе анализа треков ЛМ опорных синусоид на изображении амплитудного спектра, которая затем и применяется для синтеза нового речеподобного сигнала.

Как уже отмечалось, синтез нового РПС может проходить либо в блоке обратного БПФ с перекрываемым взвешиванием и суммированием, либо в гребенке синусоидальных генераторов (рис. 4), выход каждого звена которой модулируется амплитудой на уточненной частоте найденной опорной синусоиды и добавляется к другим найденным базовым синусоидальным волнам, чтобы сформировать окончательный вывод речи или звука, синтезированных по заданному изображению спектрограммы, которое в принципе может быть произвольного содержания.

Поскольку фазовая скорость пиков оценивается непосредственно по амплитудному спектру, а фазы непиковых бинов просчитываются через фазы пиков[17] или обнуляются, метод позволяет эффективно преобразовать спектрограмму, содержащую только величины амплитуд ЛМ, их частоту и фазу, в РПС

16 R. McAulay and T. Quatieri, Speech analysis/Synthesis based on a sinusoidal representation, in IEEE Transactions on Acoustics, Speech, and Signal Processing. V. 34, no. 4. P. 744–754, August 1986.

17 Beauregard, Gerald Harish, Mithila Wyse, Lonce Single Pass Spectrogram Inversion 2015/07/01, pp. 427–431. DOI - 10.1109/ICDSP.2015.7251907

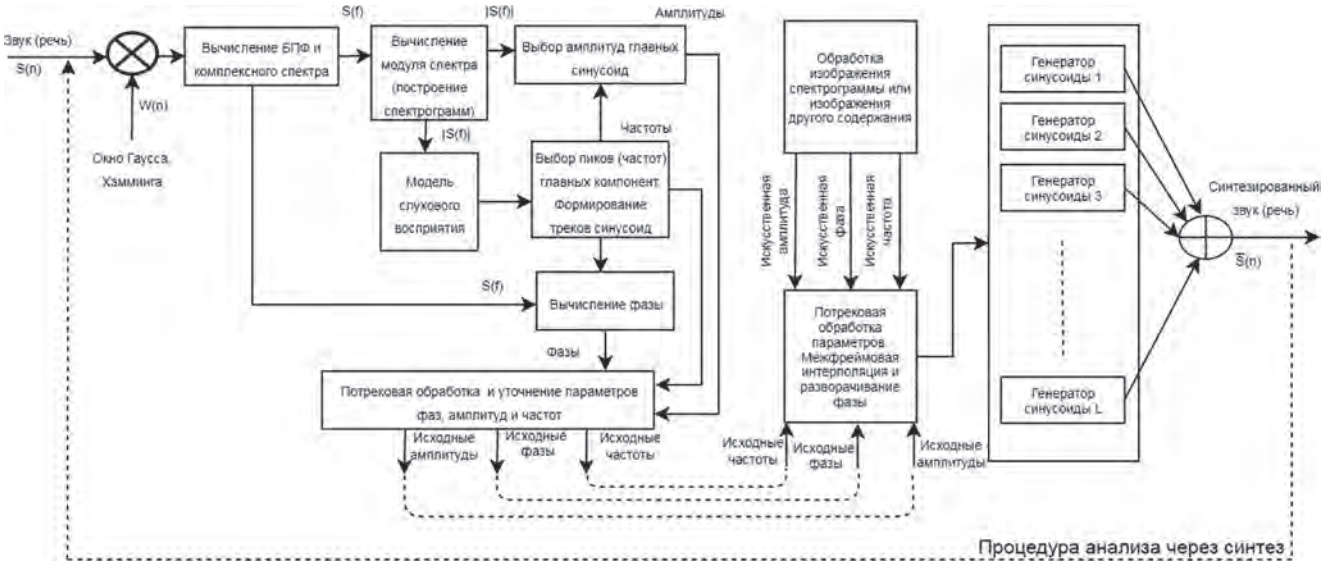


Рис. 4. Детализированная система образного анализа-синтеза акустического и речевого сигнала на основе синусоидальной узкополосной модели с учетом свойств слуха

сигнал во временной области за один детерминированный проход. То есть режим реального времени обработки в автономных системах ЗРИ вполне возможен и реализуем.

Оценка параметров локальных максимумов (пиковых бинов)

Метод однопроходной инверсии спектрограмм (ОИС) начинается с получения и обработки векторов текущего и предыдущего мгновенных амплитудных спектров (срезов), которые представляют собой результат быстрого преобразования Фурье (БПФ) над перекрывающимися и взвешиваемыми окном кадрами фактического или искусственного сигнала во временной области. Выбираемые последовательно кадры взвешиваются с помощью усеченного окна Гаусса размером σ = 0.17, базой БПФ N = 1024 и шагом анализа (перекрытия) S =50 отсчётов, что составляет 6,25 мс при Fd = 8000 Гц и длине кадра L = 1024.

Уточнение позиций (частот) локальных максимумов

Получаемые в процессе КПФ и последующей обработки спектрограмм параметры ЛМ опорных базовых синусоид, по которым впоследствии будет идти реконструкция сигнала с заданными спектральными свойствами, могут быть достаточно «загрублены» из-за некорректного выбора разрешения (базы) БПФ, вида окна и скорости трансформации спектра.

Для определения истинных значений параметров ЛМ, а именно $\{A_i;\ \omega_i;\ \varphi_{0i}\}$ производятся следующие уточняющие действия[18].

На текущем спектральном срезе определяются локальные максимумы (ЛМ) или пики при сравнении величины каждого бина j с соседями j+1 и j−1. Таким образом, если

$$|X(mS,\omega_j)|>|X(mS,\omega_{j-1})|....|X(mS,\omega_j)|>|X(mS,\omega_{j+1})|, \quad (7)$$

тогда позиция j на спектральном срезе считается пиком с амплитудой $|X(mS,\omega_j)|$.

Здесь m — это индекс времени, а

$$\omega_j = \frac{2\pi j}{N} \quad (8)$$

это частота бина j и N – база преобразования Фурье.

Далее для простоты изложения будем использовать следующие греческие буквы для описания этих параметров пикового бина и его соседей:

$$\alpha = |X(mS,\omega_{j-1})|,\ \beta = |X(mS,\omega_j)| \text{ и } \gamma = |X(mS,\omega_{j+1})| \quad (9)$$

Затем производится квадратичная интерполяция для определения истинного положения пика ЛМ на срезе, основанная на обработке позиций пикового бина и его соседей (всего три точки) с использованием формулы:

$$p = 0.5\,\frac{\alpha-\gamma}{\alpha-2\beta+\gamma} \quad (10)$$

Величина p принимает значения в диапазоне [–0.5, 0.5] и представляет собой отклонение позиции истинного пика ЛМ от пикового бина, как показано на рисунке 5. Это важно, так как каждый истинный пик соответствует синусоиде, частота которой не обязательно точно совпадает с центральной частотой БПФ бина, воспринимаемого изначально в качестве ЛМ. Рассматриваемая интерполяция дает оценку истинной пиковой частоты ЛМ.

18 Beauregard, Gerald Harish, Mithila Wyse, Lonce Single Pass Spectrogram Inversion 2015/07/01, pp. 427–431. DOI – 10.1109/ICDSP.2015.7251907

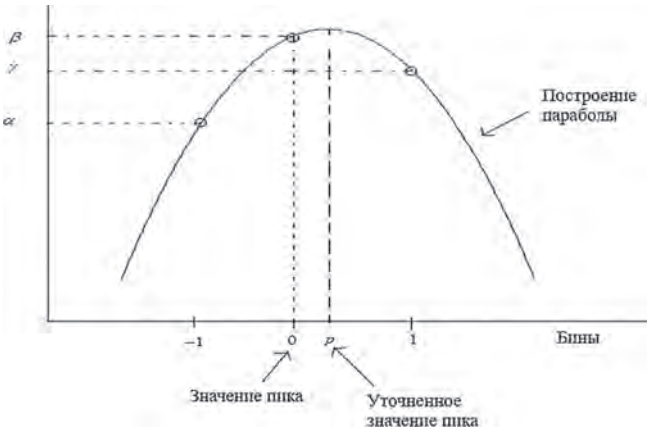


Рис. 5. Оценка значения уточненной амплитуды и частоты истинного пика с использованием квадратичной интерполяции ЛМ и позиций соседних бинов

Частота найденного истинного пика рассчитывается с использованием формулы (8), где j позиция пикового бина, а значение p рассчитывается, как в (10):

$$\omega_j = \frac{2\pi(j+p)}{N} \quad (11)$$

где ω_j — это скорректированная фазовая скорость, связанная с пиковым бином.

Если знаменатель в (10) равен 0, то истинная позиция ЛМ точно совпадает с частотой пикового бина.

Уточнение искусственной фазы ЛМ

Опираясь на результаты[19], создаем и используем фазовый аккумулятор, который хранит и пересчитывает значения фазы, которые должны использоваться для пикового бина j в текущем кадре (срезе) m, относительно фазы его предыдущего состояния. Тогда

$$\phi_{m,j} = \phi_{(m-1),j} + S\omega_j, \quad (12)$$

где S – шаг синтеза, который в большинстве приложений совпадает с размером шага анализа.

Фазовый аккумулятор каждый раз при синтезе нового среза обновляется в соответствии с формулой (11) только для пиковых бинов при формировании фаз ЛМ для нового текущего среза и сдвижке его прежних значений в предыдущий спектральный срез.

Теперь, когда фаза на пиках определена, фазы в оставшихся бинах могут быть рассчитаны в зависимости от значений признака P.

Здесь используется альтернативная стратегия π-фаз, которая может развиваться по двум сценариям: $P < 0$ и $P \geq 0$. В любом из них, два соседних слева и справа к пиковому бина будут принимать его фазу, сдвинутую на π[20].

19 Дворянкин С. В. Речевая подпись. М.: РИО-МТУСИ. 2003. – 184 с.

20 Beauregard, Gerald Harish, Mithila Wyse, Lonce Single Pass Spectrogram Inversion 2015/07/01, pp 427 – 431. DO - 10.1109/ICDSP.2015.7251907

Определение частотных позиций ЛМ, а также начальных фаз и приращения фаз в соответствии с треками движения ЛМ синусоидальных составляющих на спектрограмме сигнала использовались для синтеза нового РС и сравнения его спектрограммы с исходной по метрике Минковского.

Для реализации режима реального времени (РМВ) в условиях экономии вычислительного ресурса, предложено на своих прежних и уточненных позициях в спектральных срезах оставлять ЛМ с определенной по указанным правилам искусственной или оригинальной фазой. Все остальные точки спектра (непиковые бины) предлагается обнулять.

Варианты однопроходной инверсии спектрограмм в РМВ

Теперь, когда были рассчитаны фазы для каждого бина текущего спектрального среза, они могут быть объединены с частотными компонентами амплитудного спектра, формируя всю информацию, необходимую для реконструкции сигнала во временной области.

На последнем «вокодерном» шаге ОАС для получения реального звукового сигнала с заданными свойствами применяется либо гребенка синусоидальных генераторов (описана ранее рис. 4 для уточненных ЛМ), либо вычисляется обратное быстрое преобразование Фурье (для пиковых бинов) с взвешиванием результата усеченным окном Гаусса, чтобы получить выходные кадры, которые затем перекрываются и суммируются.

Блок-схема предлагаемого алгоритма инверсии спектрограмм реального времени, выполняемого с использованием обратного БПФ (ОБПФ) последовательно для каждого текущего среза с ЛМ, показана на рисунке 6.

Показанная на рис. 6 схема ОИС была спроектирована для работы в режиме РМВ и может входить составной частью в блок синтеза системы ОАС на рис. 4.

Под РМВ здесь понимается возможность проведения всех операций по построению одного спектрального среза, его обработки и синтеза по нему фрейма нового сигнал за промежуток времени равный интервалу между текущим и предыдущим спектральными срезами.

Рис. 6. Блок-схема алгоритма однопроходной инверсии спектрограмм с ОБПФ

Выбор метрики оценки качества синтезированных аудио сигналов

Чтобы оценить эффективность алгоритма, вычислялась амплитудная спектрограмма известного аудиосигнала во временной области, при необходимости с сохранением оригинальных фазовых значений.

Затем запускался алгоритм ОИС, чтобы получить значения искусственной фазы для генерации нового РС и получения от него новой спектрограммы, которая сравнивалась с оригиналом, используя известную меру отношения сигнала к ошибке (SER)[21] или другую меру.

В качестве исходной спектрограммы в экспериментах выступали также фотографии произвольного содержания, в частности фото лиц и предметов. Понятно, что исходные фазы здесь отсутствуют, и в процессе ОИС требуется присоединять искусственную фазу к каждому ЛМ.

$$SER=10\log\frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}X(mS,\omega)^2d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}[|X(mS,\omega)|-|X'(mS,\omega)|]^2d\omega}. \quad (13)$$

Здесь X – это результат КПФ исходного сигнала, а X' является КПФ сигнала, реконструированного с помощью ОИС, m относится к временному индексу кадров КПФ и ω – индекс круговой частоты. Высокий SER указывает на лучшее соответствие между двумя сравниваемыми спектрами.

Для простоты вычислений в качестве критерия оценки качества предлагаемого алгоритма была выбрана метрика Миньковского (знаменатель формулы (13)), с помощью которой оценивалась степень соответствия эталонного изображения, и изображения спектрограммы синтезированного по эталону сигнала, построенного с использованием усеченной оконной функции Гаусса и базой БПФ N = 1024:

$$\varepsilon=\frac{1}{HW}\{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}|C(i,j)-\hat{C}(i,j)|^2\}^{1/2} \quad (14)$$

где H и W – высота (512 точек) и ширина изображений. Соответственно, $C(i,j)$ и $\hat{C}(i,j)$ – яркости (от 0 до 255) пикселя с координатами (i,j) на двух сравниваемых изображениях.

Здесь уже малое значение ε указывает на лучшее соответствие между двумя сравниваемыми изображениями спектрограмм.

Особенности синтеза по изображению произвольного содержания содержания с прореживанием полос

Разложение сигнала на базисные функции преобразования Фурье можно также рассматривать как частный случай Гильбертовского описания синусоидальной речевой модели, когда сигнал разбивается на N/2 гармоник, узкополосных процессов с фиксированной частотой.

Чтобы образы этих гармоник не пересекались и не искажали друг друга было предложено осуществлять предварительную сепарацию используемых исходных описаний.

ТБыла дана оценка двум различным видам синтеза (ЛМ и с прореживанием полос) в случае передачи произвольного изображения в виде спектра сигнала через канал звуковой связи. Для этого исходное изображение (Рис. 7) интерпретировалось как спектрограмма некоторого звукового сообщения, по ней производился синтез речи с искусственной фазой для ЛМ. По полученному звуковому файлу строилась его спектрограмма (Рис. 8), которая затем сравнивалась с оригиналом.

Для улучшения качества передаваемой картинки исходное изображение разбивалось на несколько полос (Рис. 9). Чтобы качество итогового изображения, получаемого после склейки полос (Рис. 10), не ухудшилось между полосами делались вертикальные пробелы, пропуски.

Оптимальным оказалось разбиение на 8 полос при базе Фурье N = 1024. При дальнейшем увеличении количества полос качество итогового изображения практически не улучшается. Нетрудно заметить, что качество итогового изображения, полученного в результате синтеза с разбиением на полосы (рис. 11), значительно лучше качества изображения, синтезированного без разбиения на полосы (рис. 8).

Следует также заметить, что качество итогового изображения будет тем лучше, чем более однородно исходное изображение.

Тестирование алгоритмов синтеза

Для тестирования использовались следующие виды аудио данных: речевые сигналы (мужской и женский голос) и фото лица и предметов, представляемые как готовые спектрограммы неких звуков произвольного содержания, как-то в виде матрицы неотрицательных чисел с высотой 512 пикселей. Исходные сигналы в режиме «моно» имели продолжительность от 2 до 8 секунд и оцифровывались с частотой дискретизации 8000 Гц с 16 бит на точку отчета. Также использовались усеченное окно Гаусса длиной в 1024 точки и 50 точечное расстояние между столбцами (шаг анализа) для всех спектрограмм аудиоданных.

В таблице 1 показаны обобщенные результаты сравнения качества различных видов синтеза РС и звука по изображениям спектрограмм в разных вариантах: с использованием всех значений спектрального среза; только по локальным максимумам; по уточненным ЛМ; по ЛМ с прореживанием полос. С использованием оригинальных и-или синтетических фазовых значений. Для сравнения изображений исходной и синтезированной спектрограмм использовалась метрика Минковского.

21 Beauregard, Gerald Harish, Mithila Wyse, Lonce Single Pass Spectrogram Inversion 2015/07/01, pp. 427–431. DOI – 10.1109/ICDSP.2015.7251907

Рис. 7. Исходное изображение перед конвертацией в звуковое сообщение

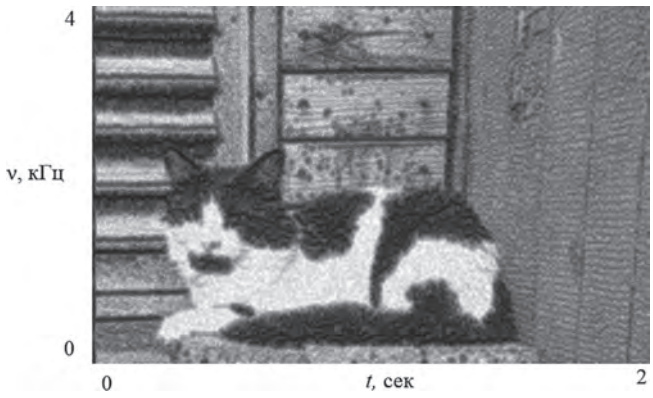


Рис. 8. Спектрограмма звукового сигнала, синтезированного только по ЛМ исходной картинки

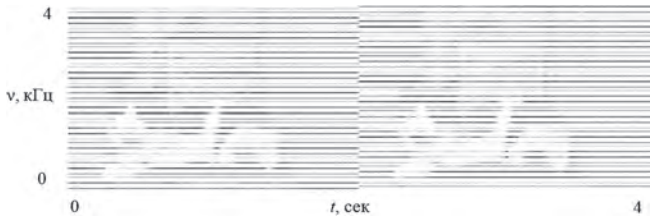


Рис. 10. Спектрограмма звукового сигнала, прореженная на полосы ЛМ (рассечение-разнесение)

Рис. 11. Восстановленное изображение после синтеза и склейки полос

По таблице 1 можно видеть, что алгоритм синтеза по всему спектральному срезу является оптимальным для точного сохранения информации об исходной спектрограмме, а синтез по локальным максимумам обладает наибольшим быстродействием (табл. 2), сохраняя при этом все идентификационные признаки и смысловое содержание речи.

В таблице 1 также отображены результаты проверки работоспособности алгоритма синтеза звука по изображению спектрограммы, в том числе для случаев, когда изображение не является сонограммой, а, например, фотопортретом.

Результаты сравнения времени работы различных видов синтеза для сигнала длительностью в 1 с представлены в таблице 2.

Из приведенных результатов тестирования видно, что синтез изображений произвольного содержания по ЛМ с прореживанием полос, оказался самым качественным, но и самым «медленным». А синтез по ЛМ для РС – самым быстрым с сохранением большинства просодических признаков.

Таким образом, под каждую задачу ЗРИ можно подобрать свой вид синтеза, удовлетворяющий заявляемым требованиям для однопроходной инверсии спектрограмм в реальном масштабе времени.

Разработанные способы и ПО генерации сигналов с требуемыми характеристиками, определяемыми заданными спектрограммами, использовались в ходе проведения экспериментов по проектам «Зрение через слух», «Речеподобные помехи, стойкие к шумоочистке», «Речевая реабилитация» и др., и показали хорошие результаты для данных видов приложений инверсии спектрограмм.

Заключение

Представлена концепция применения образного анализа-синтеза в системах защиты речевой информации (ЗРИ). В рамках предложенной концепции разработаны методы и алгоритмы однопроходной инверсии спектрограмм (ОИС), которые позволяют реализовать технологию «звук – изображение – звук», часто используемую в современных системах ЗРИ в масштабе времени близком к реальному.

Предложены различные виды синтеза звуков и речи по локальным максимумам (ЛМ) изображения исходной спектрограммы, отличающиеся по скорости исполнения и точности реализации. Наиболее быстрый – посрезный синтез по ЛМ опорных синусоид с искусственной фазой без расчета арктангенсов, наиболее точный – синтез по ЛМ базисных функций преобразования Фурье с прореживанием образа исходной спектрограммы на полосы.

Показана возможность получения оценок синтетической фазы, частоты и амплитуды ЛМ на текущем спектральном срезе, необходимых для реконструкции

Таблица 1.

Сравнение изображений спектрограмм до и после синтеза по метрике Минковского
с оригинальной и искусственной фазой для локальных максимумов спектральных срезов

Вид синтеза (предобработка спектрограммы)	Оригинальная фаза	Синтетическая фаза
Технология «Звук-Изображение-Звук - Изображение»		
по всему спектральному срезу РС, включая ЛМ и их соседей	0,003	0,009
только по локальным максимумам треков опорных синусоид	0,010	0,011
по уточненным ЛМ треков опорных синусоид	0,009	0,010
Технология «Изображение-Звук-Изображение»		
по ЛМ спектра изображения - для сонограмм - для фото лица	 информации об оригинальной фазе нет информации об оригинальной фазе нет	 0,011 0,009
по ЛМ с прореживанием полос - для сонограмм - для фото лица	 информации об оригинальной фазе нет информации об оригинальной фазе нет	 0,005 0,007

сигнала по представленной спектрограмме за один проход в РМВ.

Основным преимуществом такого метода фазовой оценки перед другими методами является простота вычислений, низкая ресурсоемкость и повышенная эффективность.

Фазовая реконструкция, выполняемая данным алгоритмом ОИС, имеет следующие преимущества перед другими методами инверсии спектрограмм:

- ✓ опирается на уточненную модель РС, представляемых в виде суммы узкополосных процессов (опорных синусоид), в качестве которых в ряде приложений могут выступать и базисные функции преобразования Фурье при соответствующем рассечении-разнесении исходной спектрограммы;
- ✓ включает в себя более точную квадратичную оценку местоположения пиков ЛМ на столбцах изображений амплитудных спектров сигнала;
- ✓ позволяет оценить фазовое приращение у соседних с пиками ЛМ на текущем спектральном срезе относительно позиций ЛМ на предыдущем;
- ✓ использует фазовый аккумулятор (стек), который работает с произвольными значениями начальных фаз для ЛМ, помогает определять фазовую функцию для трека ЛМ опорных синусоид и не требует вычислять арктангенс фазы;
- ✓ позволяет избавиться от нейросетевых вокодеров, которые сами по себе вычислительно дороги;
- ✓ может работать в режиме реального времени.

Предложенные методы ОИС также выгодно отличаются от методов инверсии спектрограмм типа

Таблица 2.

Время синтеза речевого (звукового) сообщения (мсек) длительностью в 1 с.

Вид синтеза	Оригинальная фаза	Синтетическая фаза
Синтез по всему спектральному срезу для РС	172 ± 8	182 ± 8
Синтез по локальным максимумам для РС	167 ± 8	177 ± 8
Синтез по уточненным локальным максимумам для РС	125 ± 8	134 ± 8
Синтез по ЛМ для фото лица	данных нет	177 ± 8
Синтез ИЗО по ЛМ с прореживанием полос для фото лица	данных нет	682± 8

GLA, которые требуют множества итераций частотного преобразования. Позволяют обеспечить удобные начальные фазы для итеративных методов типа GLA и производных от него, таких как FGLA, RTISI и RTISI-LA и др., что улучшает их работу и снижает количество проходов инверсии спектрограмм,

Методы ОИС обеспечивают хорошую оценку фаз с точки зрения принятой меры ошибки (мера Минковского), используемой для сравнения спектрограмм известного и синтезированного сигнала во временной области.

В рамках ОАС разработанный метод ОИС дополнительно к технологии «звук – изображение – звук» позволяет реализовать технологию «изображение – звук – изображение», формируя аудио сигнал в соответствии с любым априори заданным изображением его спектрограммы, например в виде фотопортрета.

Полученные результаты позволяют реализовать перспективные решения и расширить возможности существующих систем защиты речевой информации, сделать их более эффективными.

Литература

1. *Хорев А. А., Дворянкин С. В., Козлачков С. Б., Василевская Н. В. Анализ предельных возможностей методов шумопонижения и реконструкции речевых сигналов, маскируемых различными типами помех //Вопросы кибербезопасности. 2024. № 1 (59). С. 89–100.*
2. *Дворянкин С. В., Дворянкин Н. С., Устинов Р. А. Речеподобная помеха, стойкая к шумоочистке, как результат скремблирования защищаемой речи // Вопросы кибербезопасности. 2022. № 5 (51). С. 14–27.*
3. *Минаев В. А., Дворянкин С. В., Алюшин А. М. Методы биомаркирования защищаемых объектов // Информация и безопасность. 2023. Т. 26. № 3. С. 321–328.*
4. *Дворянкин С. В., Дворянкин Н. С., Устинов Р. А. Развитие технологий образного анализа-синтеза акустической (речевой) информации в системах управления, безопасности и связи // Безопасность информационных технологий, 2019. Т. 26, № 1. С. 64–76.*
5. *Дворянкин С. В., Зенов А. Е., Устинов Р. А., Дворянкин Н. С. Кодирование изображений спектрограмм для обеспечения переменной скорости передачи аудиоданных с сохранением качества их звучания // Безопасность информационных технологий. 2021. Т. 28. № 4. С. 22–38.*
6. *Дворянкин С. В., Уленгов С. В., Устинов Р. А., Дворянкин Н. С., Антипенко А. О. Системное моделирование речеподобных сигналов и его применение в сфере безопасности, связи и управления // Безопасность информационных технологий. 2019. Т. 26. № 4. С. 101–119.*
7. *Дворянкин С. В., Дворянкин Н. С. Средства, способы и признаки клонирования речи // Сборник статей по материалам IV Международной научно-практической конференции «Информационная безопасность: вчера, сегодня, завтра» под редакцией В. В. Арутюнова. Москва, РГГУ, 2021. С. 103–111.*
8. *Alyushin A. M., Dvoryankin S. V. Acoustic pattern recognition technology based on the Viola-Jones approach for VR and AR systems // В сборнике: Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020. Proceedings of the 11th Annual Meeting of the BICA Society. Сер. «Advances in Intelligent Systems and Computing» 2021. С. 1–8.*
9. *Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin,Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In Advances in Neural Information Processing Systems, pages 14881–14892, 2019.*
10. *Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3617–3621, 2019.*
11. *Engel J. Resnick C. Roberts A. Dieleman S. Norouzi M. Eck D., Simonyan K. Waveglow: A flow-based generative network for speech synthesis. // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019. Pp. 3617–3621.*
12. *Y. Masuyama, K. Yatabe, and Y. Oikawa, «Griffin–Lim like phase recovery via alternating direction method of multipliers», IEEE Signal Process. Lett., vol. 26, pp. 184–188, Jan. 2019.*
13. *T. Peer, S. Welker, and T. Gerkmann,«Beyond Griffin–Lim: Improved iterative phase retrieval for speech» in Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC), Sept. 2022, pp. 1–5.*
14. *Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, «Deep Griffin–Lim iteration», in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), May 2019, pp. 61–65.*
15. *«Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network», IEEE J. Sel. Top. Signal Process., vol. 15, pp. 37–50, Jan. 2021.*
16. *Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, «Fastspeech 2: Fast and high-quality end-to-end text to speech», in Proc. Int. Conf. Learn. Represent. (ICLR), May 2021.*
17. *T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, «CycleGAN-VC3: Examining and improving CycleGANVCs for mel-spectrogram conversion», in Proc. Interspeech, Oct. 2020, pp. 2017–2021.*
18. *T. Hayashi, W. C. Huang, K. Kobayashi, and T. Toda, «Nonautoregressive sequence-to-sequence voice conversion», in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), June 2021, pp. 7068–7072.*
19. *R. Prenger, R. Valle, and B. Catanzaro, «Waveglow: A flowbased generative network for speech synthesis» in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), May 2019, pp. 3617–3621.*
20. *K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Br´ebisson, Y. Bengio, and A. C. Courville, «Melgan: Generative adversarial networks for conditional waveform synthesis», in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, Dec. 2019.*
21. *J. Kong, J. Kim, and J. Bae, «Hifi-gan: Generative adversarial networks for efficient and high-fidelity speech synthesis», in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2020.*
22. *T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, «ISTFTNET: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform», in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Apr. 2022, pp. 6207–6211.*

23. J. J. Webber, C. Valentini-Botinhao, E. Williams, G. E. Henter, and S. King, «Autovocoder: Fast waveform generation from a learned speech representation using differentiable digital signal processing», arXiv:2211.06989, 2022.
24. Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, «Onoma-to-wave: Environmental sound synthesis from onomatopoeic words», APSIPA Trans. Signal, Inf. Process., vol. 11, May 2022.
25. B. D. Giorgi, M. Levy, and R. Sharp, «Mel spectrogram inversion with stable pitch», in Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR), Dec. 2022, pp. 233–239.

FAST SYNTHESIS OF AUDIO SIGNALS FROM SPECTROGRAM IMAGES IN SPEECH INFORMATION PROTECTION TASKS

Dvoryankin S. V.[22], Dvoryankin N. S.[23], Alyushin A.M.[24]

Purpose of the work: *development of methods and algorithms for spectrogram inversion: determination of the waveform of a signal using the previously known data of its amplitude spectral sweeps in the absence of phase information – for real-time generation of audio signals with specified frequency-temporal properties with their subsequent application in speech information protection systems.*

Research methods: *applied system analysis, digital spectral-temporal analysis, digital signal and image processing, image analysis of sonograms.*

Research results: *methods and algorithms of synthesis of sound and speech signals by a priori given spectrogram, realized within the framework of the concept of image analysis-synthesis, working in real time and providing good qualitative estimates of the phase of peak values of spectral slices in one fully deterministic pass, are proposed. They can be used alone or in obtaining initial phase estimates to improve the results of iterative algorithms like Griffin-Lim et al. The estimates of positions and phase of spectral peaks obtained from the processed spectrogram images are determined more accurately using quadratic interpolation, and the recalculation of the phase increment by time steps is performed in a specially introduced phase accumulator, without requiring the calculation of arctangents.*

Scientific novelty: *a new method of spectrogram inversion based on dissection-dissection of the original spectrogram image is proposed to obtain more accurate spectral descriptions of the audio signal synthesized from it, better corresponding to the original than known iterative methods of spectral inversion.*

Practical value: *a computationally efficient real-time algorithm for single-pass spectrogram inversion has been developed. The obtained results will allow to expand the capabilities of existing systems of speech information protection and to design more effective ones on the basis of the described approaches.*

Keywords: *information security, spectrogram inversion, image analysis, protection against unauthorized access, speech-like signal, sinusoidal speech model.*

References

1. Khorev A. A., Dvoryankin S. V., Kozlachkov S. B., Vasilevskaya N. V. Analiz predel'nykh vozmozhnostei metodov shumoponizheniya i rekonstruktsii rechevykh signalov, maskiruemykh razlichnymi tipami pomekh //Voprosy kiberbezopasnosti. 2024. № 1 (59). S. 89–100.
2. Dvoryankin S. V., Dvoryankin N. S., Ustinov R. A. Rechepodobnaya pomekha, stoikaya k shumoochistke, kak rezul'tat skremblirovaniya zashchishchaemoi rechi // Voprosy kiberbezopasnosti. 2022. № 5 (51). S. 14–27.
3. Minaev V. A., Dvoryankin S. V., Alyushin A. M. Metody biomarkirovaniya zashchishchaemykh ob"ektov // Informatsiya i bezopasnost'. 2023. T. 26. № 3. S. 321–328.
4. Dvoryankin S. V., Dvoryankin N. S., Ustinov R. A. Razvitie tekhnologii obraznogo analiza-sinteza akusticheskoi (rechevoi) informatsii v sistemakh upravleniya, bezopasnosti i svyazi // Bezopasnost' informatsionnykh tekhnologii, 2019. T. 26, № 1. C. 64–76.
5. Dvoryankin S. V., Zenov A. E., Ustinov R. A., Dvoryankin N. S. Kodirovanie izobrazhenii spektrogramm dlya obespecheniya peremennoi skorosti peredachi audiodannykh s sokhraneniem kachestva ikh zvuchaniya // Bezopasnost' informatsionnykh tekhnologii. 2021. T. 28. № 4. S. 22–38.
6. Dvoryankin S. V., Ulengov S. V., Ustinov R. A., Dvoryankin N. S., Antipenko A. O. Sistemnoe modelirovanie rechepodobnykh signalov i ego primenenie v sfere bezopasnosti, svyazi i upravleniya // Bezopasnost' informatsionnykh tekhnologii. 2019. T. 26. № 4. S. 101–119.
7. Dvoryankin S. V., Dvoryankin N. S. Sredstva, sposoby i priznaki klonirovaniya rechi // Sbornik statei po materialam IV Mezhdunarodnoi nauchno-prakticheskoi konferentsii «Informatsionnaya bezopasnost': vchera, segodnya, zavtrA» pod redaktsiei V. V. Arutyunova. Moskva, RGGU, 2021. S. 103–111.

22 Sergey V. Dvoryankin, Dr.Sc. (of Tech.), Professor, Professor of the Department of Strategic Information Studies, National Research Nuclear University MEPhI, Head of the Laboratory for the Protection and Processing of Audiovisual Information, Moscow State Linguistic University. Moscow. Russia. E-mail: svdvoryankin@mephi.ru. https://orcid.org/0000-000-6908-0676
23 Nikita S. Dvoryankin, postgraduate student, National Research Nuclear University MEPhI. Moscow. Russia. E-mail: nik.dvrn@gmail.com
24 Alexander M. Alyushin, Senior Lecturer, Department of Informatics and Control Processes, National Research Nuclear University MEPhI, Researcher, Laboratory of Protection and Processing of Audiovisual Information., Moscow State Linguistic University. Moscow. Russia. E-mail: alyshin@list.ru

8. *Alyushin A. M., Dvoryankin S. V. Acoustic pattern recognition technology based on the Viola-Jones approach for VR and AR systems. В сборнике: Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020. Proceedings of the 11th Annual Meeting of the BICA Society. Сер. «Advances in Intelligent Systems and Computing» 2021. С. 1–8.*
9. *Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin,Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In Advances in Neural Information Processing Systems, pages 14881–14892, 2019.*
10. *Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3617–3621, 2019.*
11. *Engel J. Resnick C. Roberts A. Dieleman S. Norouzi M. Eck D., Simonyan K. Waveglow: A flow-based generative network for speech synthesis. // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019. Pp. 3617–3621.*
12. *Y. Masuyama, K. Yatabe, and Y. Oikawa, «Griffin–Lim like phase recovery via alternating direction method of multipliers», IEEE Signal Process. Lett., vol. 26, pp. 184–188, Jan. 2019.*
13. *T. Peer, S. Welker, and T. Gerkmann,«Beyond Griffin–Lim: Improved iterative phase retrieval for speech» in Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC), Sept. 2022, pp. 1–5.*
14. *Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, «Deep Griffin–Lim iteration», in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), May 2019, pp. 61–65.*
15. *«Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network», IEEE J. Sel. Top. Signal Process., vol. 15, pp. 37–50, Jan. 2021.*
16. *Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, «Fastspeech 2: Fast and high-quality end-to-end text to speech», in Proc. Int. Conf. Learn. Represent. (ICLR), May 2021.*
17. *T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, «CycleGAN-VC3: Examining and improving CycleGANVCs for mel-spectrogram conversion», in Proc. Interspeech, Oct. 2020, pp. 2017–2021.*
18. *T. Hayashi, W. C. Huang, K. Kobayashi, and T. Toda, «Nonautoregressive sequence-to-sequence voice conversion», in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), June 2021, pp. 7068–7072.*
19. *R. Prenger, R. Valle, and B. Catanzaro, «Waveglow: A flowbased generative network for speech synthesis» in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), May 2019, pp. 3617–3621.*
20. *K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Br´ebisson, Y. Bengio, and A. C. Courville, «Melgan: Generative adversarial networks for conditional waveform synthesis», in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, Dec. 2019.*
21. *J. Kong, J. Kim, and J. Bae, «Hifi-gan: Generative adversarial networks for efficient and high-fidelity speech synthesis», in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2020.*
22. *T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, «ISTFTNET: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform», in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Apr. 2022, pp. 6207–6211.*
23. *J. J. Webber, C. Valentini-Botinhao, E. Williams, G. E. Henter, and S. King, «Autovocoder: Fast waveform generation from a learned speech representation using differentiable digital signal processing», arXiv:2211.06989, 2022.*
24. *Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, «Onoma-to-wave: Environmental sound synthesis from onomatopoeic words», APSIPA Trans. Signal, Inf. Process., vol. 11, May 2022.*
25. *B. D. Giorgi, M. Levy, and R. Sharp, «Mel spectrogram inversion with stable pitch», in Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR), Dec. 2022, pp. 233–239.*