УДК 004.032.6; 004.627

В.Ю. Герасимович, Ал.А. Петровский

ПСИХОАКУСТИЧЕСКИ МОТИВИРОВАННОЕ ПОСТРОЕНИЕ СЛОВАРЯ ЧАСТОТНО-ВРЕМЕННЫХ ФУНКЦИЙ УНИВЕРСАЛЬНОГО МАСШТАБИРУЕМОГО АУДИОКОДЕРА НА ОСНОВЕ РАЗРЕЖЕННОЙ АППРОКСИМАЦИИ

Рассматривается способ построения перцептуально-мотивированного словаря частотно-временных функций на основе оптимизированного для фрейма входного сигнала пакетного дискретного вейвлет-преобразования и использования этого способа в универсальном масштабируемом аудиокодере реального времени. Показывается актуальность данной задачи, большое внимание уделяется психоакустическому моделированию. Описываются такие алгоритмы, как разреженная аппроксимация, перцептуальная адаптация дерева декомпозиции пакетного дискретного вейвлет-преобразования, а также схемы кодирования и декодирования входного сигнала. Приводятся результаты экспериментальных исследований разрабатываемого аудиокодера. Дается его сравнение с современными схемами сжатия звуковой информации, такими как Opus и Vorbis, на базе объективной оценки качества PEAQ – ODG.

Введение

Разреженная аппроксимация сигнала является удобным инструментом для построения алгоритмов аудиокодирования, поскольку ее суть заключается в представлении входной информации минимальным количеством ненулевых компонентов (атомов). Одним из представителей данного класса алгоритмов является согласованная подгонка (СП) [1].

В существующих подходах к аудиокодированию СП может применяться как одна из ступеней алгоритма сжатия [2, 3]. Такие варианты кодирования разделяют входной аудиосигнал на три компонента: гармонический (синусоидальный), переходный (транзиентный) и шумовой. В работе [3] синусоидальные и транзиентные компоненты моделируются на основе СП, однако для каждой из частей применяется свой словарь функций, шумовая (остаточная) часть сигнала параметризируется с помощью линейного предсказания. Суть подхода [2] состоит также в выделении трех компонентов из входного фрейма сигнала, однако СП используется только для моделирования переходных компонентов. Данные трехкомпонентные подходы, с одной стороны, позволяют получить подробную и надежную модель сигнала, но с другой – требуют трех разных алгоритмов обработки компонентов, а это обусловливает высокую вычислительную сложность всего метода. Существуют также варианты построения алгоритма кодирования звука с использованием СП для параметризации аудиосигнала без разделения его на определенные компоненты. К примеру, в работе [4] показан вариант моделирования аудиосигнала на основе СП с использованием словарей частотно-временных функций Габора. В данном подходе перцептуальная модель используется для уменьшения итогового количества атомов, но, в силу того что эффект маскирования применяется после процедуры СП, длительность работы алгоритма разреженной аппроксимации может быть довольно большой. В работе [5] также применяется метод СП для моделирования сигнала, где словарь функций строится на основе модифицированного дискретно-косинусного преобразования (МДКП) с различной длиной базисных функций. В данном подходе, как и в предыдущем, в процессе СП психоакустическое моделирование не применяется (применяется контроль эффекта пре-эхо).

Во всех подходах аудиокодирования на основе СП самыми важными задачами являются построение оптимального словаря частотно-временных функций и определение признака остановки работы алгоритма аппроксимации. Исследования, описываемые в данной статье, посвящены решению этих задач и построению аудиокодера на основе разрабатываемой модели. Здесь представлена модель универсального масштабируемого аудиокодера реального времени, инвариантного к звуковому информационному наполнению обрабатываемого аудиосигнала. Основой алгоритма сжатия является аппарат СП со словарями частотно-временных функций,

который позволяет осуществить операцию разреженной аппроксимации входного сигнала. Словарь частотно-временных функций строится на основе пакетного дискретного вейвлет-преобразования (ПДВП), которое динамически оптимизируется для фрейма входного сигнала с помощью психоакустического критерия [2]. Важной особенностью алгоритма является тот факт, что словарь функций формируется на базе входного обрабатываемого сигнала. Возможность масштабируемости битового потока совместно с работой в реальном масштабе времени позволяет использовать разрабатываемый алгоритм в таких областях, как системы передачи речи и звука по цифровым коммуникационным каналам, например Voice Over Internet Protocol (VoIP), Voice Over LTE (VoLTE); в сервисах потокового мультимедиа (Streaming Media) и цифрового радиовещания (Digital Audio Broadcasting, DAB).

1. Определение зависимости структуры словаря частотно-временных функций от психоакустического критерия

ПДВП – это преобразование, позволяющее получить неравномерный частотно-временной план обрабатываемого сигнала. Такая возможность существует благодаря самой сути ПДВП – итеративной декомпозиции пространства сигнала на НЧ- и ВЧ-составляющие (двоичное дерево декомпозиции) [6]. Исходя из этого, имеется возможность выбрать из общего множества такой вариант декомпозиции (т. е. частотно-временного плана), который будет эффективно описывать входной сигнал в вейвлет-области. Пример построения неравномерного частотно-временного плана в соответствии с произвольным двоичным деревом декомпозиции показан на рис. 1, где каждый квадрат представляет собой область с временным разрешением t и частотным f. Частотное разрешение увеличивается в два раза с ростом дерева вглубь на каждый новый уровень l. Временное разрешение определяется в зависимости от номера уровня l как 2^l. Согласно принципу неопределенности [7] невозможно одновременно получить высокое частотное и временное разрешение, т. е. при увеличении высоты квадрата будет уменьшаться его ширина, и наоборот. В соответствии с этим необходимо адаптировать частотно-временной план каждого фрейма согласно перцептуальному критерию как самому оптимальному с точки зрения восприятия звука слуховой системой человека.

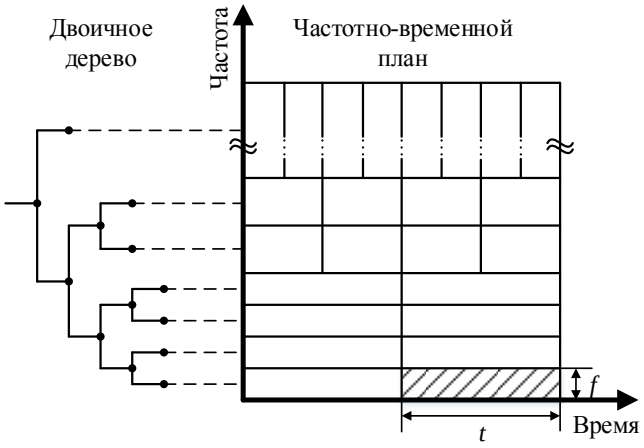


Рис. 1. Частотно-временной план для двоичного дерева вейвлет-пакета

Фактически частотно-временной план представляет собой графическое отображение дерева декомпозиции ПДВП E, которое, в свою очередь, является отражением словаря атомов D для алгоритма СП. Следовательно, определение оптимального словаря сводится к поиску такого дерева E из множества допустимых, которое позволит выделить (и, соответственно, даст алгоритму СП учесть) все особенности обрабатываемого фрейма сигнала. Поскольку входными данными являются аудиосигналы, в алгоритме определения словаря необходимо использовать закономерности психоакустики для того, чтобы учесть особенности восприятия звука человеком: перцептуальная модель позволяет определять те компоненты сигнала, которые будут оказывать максимальное влияние на восприятие звука слуховой системой человека.

В настоящей работе используется определение порогов в частотной (simultaneous) и временной (temporal) вейвлет-области [8, 9]. Общая схема алгоритма расчета частотного порога маскирования ($T_{l,n}$) показана на рис. 2.

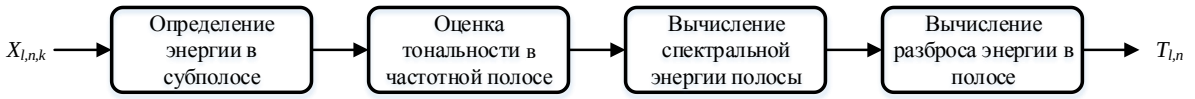


Рис. 2. Расчет порога $T_{l,n}$

На вход алгоритма поступают коэффициенты ПДВП $X_{l,n,k}$, где k – номер коэффициента на уровне l в узле n дерева декомпозиции E_j. Определение энергии в субполосе осуществляется с нормировочным делителем $\sqrt{2}$ для каждого уровня дерева E_j, так как данное значение имеют коэффициенты усиления фильтров, соответствующих семейству Добеши 20 для ПДВП [10]. Коэффициент тональности показывает тип маскирования, которое происходит в анализируемой полосе и рассчитывается на основе индексов маскирования тоном шума и шумом шума с учетом значения меры спектральной пологости [11, 12]. Именно значение этой меры показывает тональную либо шумоподобную природу сигнала: значение, равное –60 дБ, соответствует полностью тональному сигналу в полосе, 0 дБ – полностью шумоподобному [11, 13]. Спектральная энергия вычисляется с учетом энергии в субполосе и определенного ранее коэффициента тональности. Далее свертка спектральной энергии полосы с функцией разброса [9] позволяет сделать оценку частотного порога маскирования в анализируемой субполосе.

Схема расчета временного маскера ($F_{l,n}$) изображена на рис. 3.

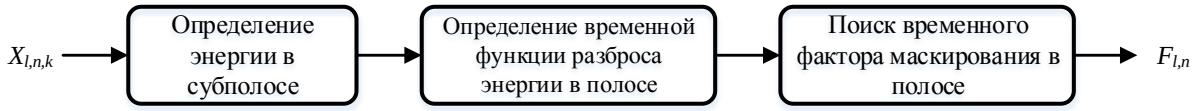


Рис. 3. Схема оценки временного порога маскирования

Как и в случае расчета частотного порога маскирования $T_{l,n}$, при определении энергии производится нормировка входных коэффициентов. Временная функция разброса энергии в полосе определяется как свертка энергии и функции разброса (с параметрами, соответствующими рассчитываемому временному порогу). Временной фактор маскирования – это определение того, присутствует в анализируемой субполосе временной маскер или нет. Данный поиск осуществляется путем сравнения временной функции разброса и энергии в полосе. В случае если значение функции разброса больше либо равно энергии сигнала в полосе, в данной субполосе присутствует временное маскирование, а его значение равно значению функции.

Глобальный порог маскирования, который учитывает как частотный, так и временной маскеры, вычисляется следующим образом:

$$M_{l,n} = T_{l,n} \cdot F_{l,n},$$

$$Gt_{l,n} = \max(M_{l,n}, ATH_{l,n}),$$

где $M_{l,n}$ – частотно-временной маскирующий порог, $ATH_{l,n}$ – порог абсолютной слышимости, $G_{l,n}$ – глобальный порог маскирования. Пример глобального порога маскирования для одного фрейма входного сигнала представлен на рис. 4.

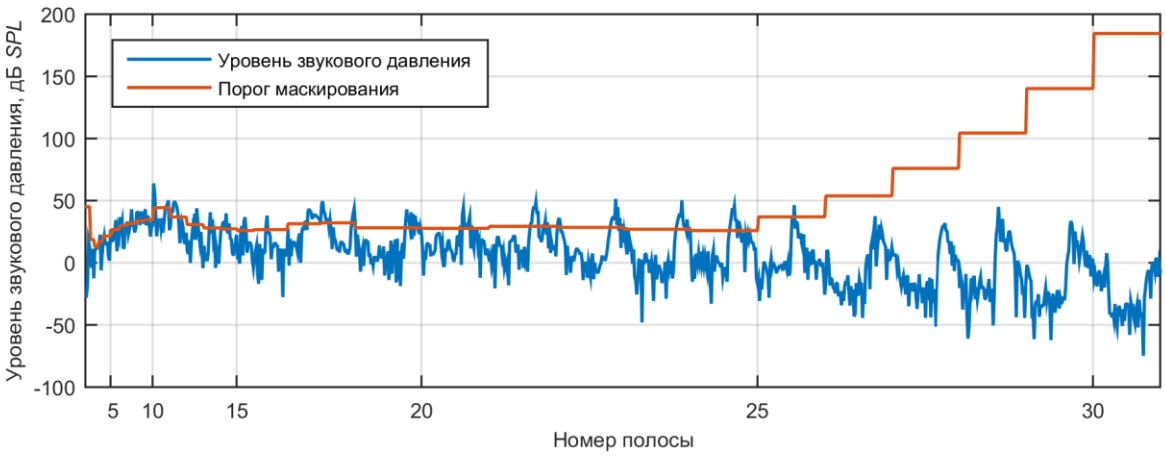


Рис. 4. Порог маскирования для одного фрейма входного сигнала

Эффект маскирования, как показано, например, в [13], может простираться во времени как до непосредственного появления маскера, так и после него. Такое свойство порога маскирования получило название пре- и постмаскирования. В то время как эффект премаскирования длится несколько миллисекунд и может находиться в пределах анализируемого фрейма, постмаскирование может длиться от 50 до 300 мс, т. е. маскер текущего фрейма будет оказывать влияние на значение порога маскирования последующих анализируемых участков аудиосигнала. Для того чтобы учесть данный эффект, в разрабатываемом аудиокодере было реализовано накопление предыстории о временных маскерах за J смежных фреймов и произведен расчет текущего как их усреднение.

Алгоритм расчета порога маскирования с учетом эффекта постмаскирования

ПОВТОРЯТЬ $\forall$ *frameN*

ЕСЛИ *тип порога* == «частотный» **ТО**

ПОВТОРЯТЬ $\forall$ частотные полосы (z):

Спектральная энергия полосы: $A(z) = \sum_{k=0}^{K-1} X_{z,k}^2$.

Мера спектральной пологости:

$$SFM(z) = \left(\prod_{k=0}^{K-1} X_{z,k}^2 \right)^{1/K} \Big/ \left(\frac{1}{K} \sum_{k=0}^{K-1} X_{z,k}^2 \right).$$

Оценка тонального коэффициента: $\eta = \min(|\, SFM(z)/60\,|, 1)$.

Расчет значения тональности маскеров:

$a(z) = \eta a_{tmn}(z) + (1-\eta) a_{nmn}(z)$;

$a_{tmn}(z) = -0{,}275z - 15{,}025z$;

$a_{nmn}(z) = -25$.

Спектр энергии полосы с учетом тональности:

$D(z) = 10 \log_{10}(A(z) \cdot 10^{a(z)/10})$.

Функция разброса энергии:

$$B(z) = a + \frac{1}{2} \cdot (v+u) \cdot (z+c) - \frac{1}{2} \cdot (v-u) \sqrt{d + (z+c)^2}\,.$$

Вычисление порога:

$$T_{l,n} = 10 \log_{10} \left(\frac{1}{K} \sum_{k=1}^{K} 10^{\frac{D(z)}{10}} \cdot 10^{\frac{B(z-K)}{10}} \right).$$

КОНЕЦ

ЕСЛИ *тип порога* == «временной» **ТО**

ПОВТОРЯТЬ $\forall$ частотные полосы (z):

Расчет энергии коэффициентов в полосе: $E_z(k) = X_{z,k}^2$.

Расчет временной функции разброса энергии $B(z)$.

Вычисление порога: $F_{l,n}(k) = \frac{1}{K}\sum_{k=0}^{K} E_z(k) \cdot 10^{\frac{B(K-k)}{10}}$.

КОНЕЦ

ЕСЛИ mod(*frameN*, *J*) != 1 **ТО** выключить алгоритм роста дерева E_{frameN}.

ПОВТОРЯТЬ $\forall\,(l,n):\quad F_{l,n}(frameN) = \left(F_{l,n}(frameN) + F_{l,n}(frameN-1)\right)/2$.

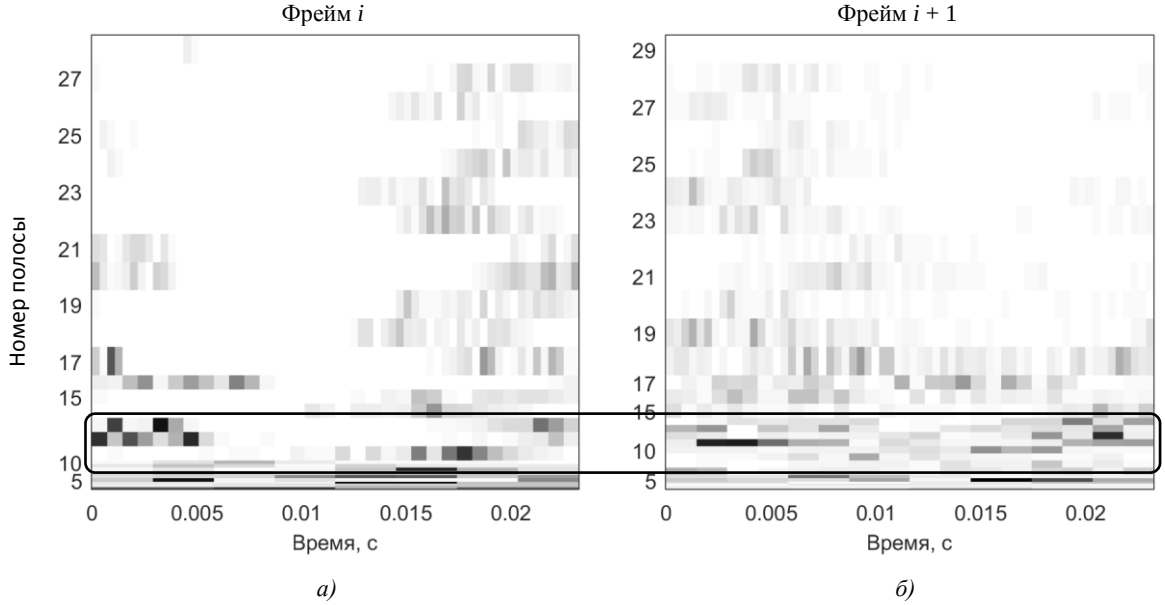
ИНАЧЕ включить алгоритм роста дерева E_{frameN}.

ЕСЛИ *тип порога* == «глобальный» **ТО** $Gt_{l,n} = \max\left(T_{l,n} \cdot \max\left(F_{l,n}, 1\right), ATH_{l,n}\right)$.

КОНЕЦ

В приведенном выше алгоритме *frameN* – номер текущего анализируемого фрейма, *J* – количество фреймов, в течение которых будет анализироваться предыстория при расчете эффекта постмаскирования. Для каждого фрейма обрабатываемого сигнала осуществляется расчет трех типов порогов маскирования: частотного порога $T_{l,n}$, временного маскера $F_{l,n}$ и глобального частотно-временного порога $Gt_{l,n}$. Операция mod (делимое, делитель) определяет остаток после деления. Процедура учета эффекта постмаскирования внедряется на стадии расчета временного порога. На первом шаге определяется, кратен ли номер текущего фрейма значению *J*, поскольку для накопления информации о $F_{l,n}$ необходимо, чтобы структура дерева декомпозиции у *J* смежных фреймов совпадала. Следовательно, алгоритм роста дерева ПДВП должен включаться только каждый (*J*+1)-й фрейм обработки. В случае если номер текущего фрейма не равен (*J*+1), производится усреднение $F_{l,n}$ со значением временного маскера из предыдущего фрейма, в противном случае требуется перерасчет новой структуры дерева декомпозиции ПДВП.

На рис. 5 видно, что субполосные разбиения двух смежных фреймов несколько отличаются. Так, например, нижние прямоугольники выделенной области частотно-временного плана на рис. 5, *а* имеют большую протяженность по временной оси и при этом меньшую по частотной (частотная шкала размечена в полосах дерева декомпозиции ПДВП) по сравнению с планом рис. 5, *б*. Это означает, что в данной области потребовалось большее частотное разрешение. Затем картина меняется, поскольку в частотно-временном плане на рис. 5, *а* прямоугольники сверху выделенной области имеют большее временное разрешение, нежели на плане рис. 5, *б*, что и отражено в их геометрических габаритах.



а) *б)*

Рис. 5. Частотно-временные планы двух смежных фреймов

На рис. 6 показана визуализация изменения порога маскирования для трех последовательных фреймов входного сигнала. Ось, отражающая частотное направление, размечена в полосах дерева декомпозиции ПДВП, которые для детализации и наглядности графически выполнены однообразными.

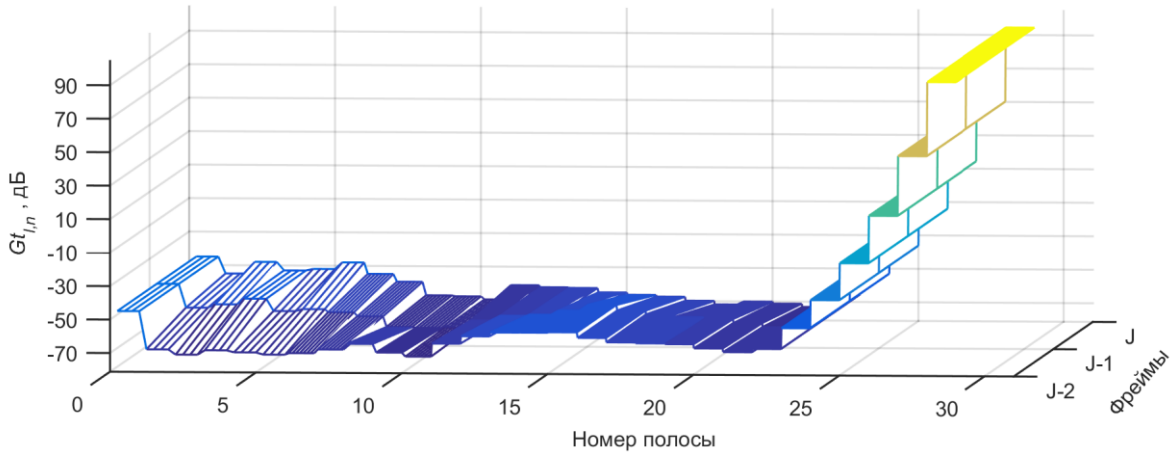


Рис. 6. Глобальный порог маскирования трех последовательных фреймов сигнала с учетом эффекта постмаскирования

На основе рассчитанной психоакустической модели для анализируемого фрейма входного сигнала можно оценить перцептуальную значимость вейвлет-коэффициентов каждого уровня ПДВП, а значит, и оптимальный частотно-временной план.

2. Разреженная аппроксимация с перцептуально-мотивированным словарем частотно-временных функций на основе ПДВП

2.1. Алгоритм СП со словарем частотно-временных функций на базе адаптивного ПДВП

СП представляет собой жадный алгоритм, суть которого заключается в отображении входного сигнала ($x(i)$) на избыточный словарь (D) частотно-временных функций (g_γ), также называемых атомами. Избыточность словаря означает, что в нем содержится намного больше элементов, чем минимально необходимое базисное количество для представления сигнала [1, 14, 15]. Данный алгоритм является итеративным. На каждом его шаге происходит поиск атома из словаря, у которого будет максимальная корреляция с фреймом моделируемого сигнала, т. е. максимальное скалярное произведение выбираемого атома и моделируемого сигнала.

Одной из необходимых характеристик разрабатываемого аудиокодера является инвариантность к информационному наполнению входных аудиоданных. Следовательно, построение словаря D частотно-временных функций должно быть адаптировано к фрейму обрабатываемого сигнала. Учитывая этот факт, оптимальным будет построение словаря на основе входной обрабатываемой информации. Такое решение можно получить, используя ПДВП [16]. Применительно к данному преобразованию словарь атомов представляет собой семейство ортонормированных базисных функций, которые имеют высокую частотную и временную локализацию. Поскольку алгоритм СП является ресурсоемким, используется адаптированное под перцептуальную специфику фрейма входного сигнала дерево декомпозиции ПДВП [8, 16, 17], что находит отражение в структуре частотно-временного плана преобразования.

Рост дерева декомпозиции ПДВП в представленной работе происходит динамически в результате оценки двух стоимостных функций [18]: перцептуальной энтропии (Perceptual Entropy, PE), значение которой соответствует локальной перцептивной информации каждого узла текущего дерева, и временной энтропии (Wavelet Time Entropy, WTE), которая показывает информативность вейвлет-коэффициентов каждого уровня дерева ПДВП:

$$\Delta_{l,n} = \sqrt{12 \cdot Gt_{l,n} / K_{l,n}}; \tag{1}$$

$$PE_{l,n} = \sum_{k=1}^{K_{l,n}-1} \log_2 \left(2 \left[n\mathrm{int} \left(\left| X_{l,n,k} \right| / \Delta_{l,n} \right) \right] + 1 \right); \tag{2}$$

$$WTE_{E_j} = - \sum_{\forall (l,n) \in E_j} \sum_{k} \frac{\left| X_{l,n,k} \right|}{\sum_{\forall (l,n) \in E_j} \left| X_{l,n,k} \right|} \ln \left(\frac{\left| X_{l,n,k} \right|}{\sum_{\forall (l,n) \in E_j} \left| X_{l,n,k} \right|} \right), \tag{3}$$

где $\Delta_{l,n}$ – шаг квантования, $K_{l,n}$ – количество вейвлет-коэффициентов X в узле (l, n) двоичного дерева декомпозиции ПДВП E_j, $Gt_{l,n}$ – глобальный порог маскирования.

Алгоритм расчета оптимального дерева декомпозиции ПДВП

Инициализация: $l = 0$, $n = 0$, $X_{l,n,k} = x(i)$, $j = 0$, СТОП = 0.
ПОВТОРЯТЬ:
Декомпозиция: $X_{l,n,k}$ деревом E_{j+1}.
Вычисление: WTE_{E_j}, WTE_{E_j+1}.
ЕСЛИ $WTE_{E_j} \geq WTE_{E_{j+1}}$ И $E_{j+1} \leq E_{CB}$ **ТО** $E_j = E_{j+1}$, $j = j + 1$.
ИНАЧЕ СТОП = 1.
Вычисление: PE $\forall$ узла E_j.
ЕСЛИ $PE_{l,n} \geq PE_{l+1,2n} + PE_{l+1,2n+1}$ **ТО** расщепление (l, n).
ИНАЧЕ не расщеплять (l, n).
ПОКА СТОП != 1.

Как было сказано выше, алгоритм определения дерева декомпозиции ПДВП для текущего фрейма анализируемого сигнала основывается на оценке двух стоимостных функций: WTE и PE. На каждой итерации процедуры вычисления роста дерева ПДВП оцениваются данные функции для текущей структуры E_j и сравниваются с таковыми для предыдущей структуры. В случае если происходит увеличение значения временной энтропии или же конфигурация дерева стала больше либо равна максимально возможной (E_{CB}), алгоритм останавливает свою работу. В противном случае производится сравнение перцептуальной энтропии для того, чтобы оценить, какие узлы необходимо расщепить, а какие останутся в итоговом дереве декомпозиции ПДВП. Рост дерева, оценка стоимостных функций, а также реализация психоакустической модели происходят «на ходу», сверху вниз – от корневого узла дерева к оконечным (терминальным) узлам, что позволяет минимизировать вычислительные затраты алгоритма оптимизации дерева ПДВП и расчета порогов маскирования.

После расчета оптимального для текущего входного фрейма сигнала дерева декомпозиции E_j ПДВП включается алгоритм согласованной подгонки, который представлен ниже.

Алгоритм СП с перцептуально-мотивированным словарем частотно-временных функций

Инициализация: $r_1 = x(i)$, $m = 0$, $Gt_{l,n}^m$, СТОП = 0.

ПОВТОРЯТЬ $\forall$ (l, n, k):

Выбрать: $X_{l,n,k}^* \in X_{l,n,k}^m$ с максимальным весом.

Поиск: a_m в $X_{l,n,k}^*$, с $\max \left(\frac{X_{l,n,k}^{*\ 2}}{Gt_{l,n}^m} \right)$.

Сохранить текущие (l, n, k) для a_m.

Синтезировать g_γ на основе a_m и ПДВП$^{-1}$.

Вычислить сигнал-остаток: $r_{m+1} = r_m - a_m g_\gamma$,

$m = m + 1$.

Пересчитать $Gt_{l,n}^m$.

Декомпозиция сигнала r_{m+1}.
ЕСЛИ критерий остановки == ИСТИНА **ТО** СТОП = 1
ПОКА СТОП != 1.

Суть алгоритма СП состоит в итеративном выборе атомов из словаря, который в данном случае представлен в виде декомпозиции входного сигнала оптимизированным деревом ПДВП. На первом шаге отбора происходит определение коэффициентов с максимальным весом возбуждения в каждой частотной полосе. Затем на базе выбранных данных необходимо выполнить поиск такого коэффициента, который бы максимизировал соответствие скалограммы для оригинального и моделируемого входного фрейма сигнала. В данном случае это эквивалентно поиску такого коэффициента, который обладает максимальным соотношением «сигнал – порог маскирования». В качестве порога маскирования используется глобальный порог $Gt_{l,n}$, который учитывает как частотное маскирование, так и временное. После того как атом будет найден на текущей итерации работы алгоритма СП, необходимо вычесть его вклад в формирование входного сигнала, затем пересчитать глобальный порог маскирования для сигнала-остатка и осуществить его декомпозицию для работы следующей итерации. После достижения критерия остановки алгоритм завершит работу. Выбор критерия остановки алгоритма является крайне важной задачей при применении СП. Для использования СП в алгоритме кодирования аудиосигналов необходимо определить сбалансированную стратегию, согласно которой процедура СП будет завершать работу. В разрабатываемом аудиокодере есть возможность устанавливать критерий остановки по количеству выбранных атомов и по энергии остаточного сигнала. Также есть возможность устанавливать перцептуальный критерий, т. е. проводить анализ остаточного сигнала на предмет того, насколько важная с точки зрения восприятия информация осталась в данном сигнале, и на базе этой информации делать вывод об остановке работы алгоритма СП.

2.2. Схемы кодирования и декодирования на основе разреженной аппроксимации

Схема процесса кодирования универсального масштабируемого аудиокодера на основе разреженной аппроксимации изображена на рис. 7.

Рис. 7. Схема процесса кодирования входного сигнала

Алгоритм кодирования состоит из следующих основных шагов. Во время инициализации входной сигнал разделяется на фреймы длительностью 1024 отсчета (для аудиосигналов с частотой дискретизации 44,1 кГц) с перекрытием между соседними фреймами в 1/8 длины фрейма; на шаге «анализ ПДВП» выполняется построение адаптированного к моделируемому сигналу дерева ПДВП E_j и декомпозиция входного фрейма, а также расчет порогов маскирования $Gt_{l,n}$; на шаге «согласованная подгонка» происходит выбор наиболее перцептуально важных для восприятия атомов и расчет сигнала-остатка для текущей итерации m алгоритма СП. Отобранные данные – атомы и их позиции – передаются на шаг «кодирование и квантование» для формирования результирующего потока битов, который будет передан декодеру. Шаг «кодирование и квантование» подробнее описан в подразд. 2.3.

Структура процесса декодирования данного аудиокодера показана на рис. 8.

Рис. 8. Схема алгоритма декодирования сигнала

На шаге «восстановление параметров» алгоритма декодирования входного потока битов производится деквантование и декодирование данных. Затем на шаге «размещение атомов» по координатам *l*, *n*, *k* все отобранные атомы устанавливаются в соответствующие узлы дерева реконструкции ПДВП. Благодаря этому обратный ПДВП осуществляется по максимальному дереву, предусмотренному алгоритмом, без необходимости вычисления его структуры. На шаге «реконструкция сигнала» производится синтез выходного фрейма аудиосигнала с помощью обратного ПДВП.

2.3. Кодирование и квантование атомов

Схема кодирования состоит из двух основных блоков: квантования и энтропийного кодирования. Параметрами, которые необходимо закодировать и заквантовать, являются атомы, а также их позиции в дереве ПДВП. Пример части множества параметров для произвольного фрейма входного сигнала представлен в табл. 1.

Таблица 1

Пример пяти параметров произвольного фрейма входного сигнала

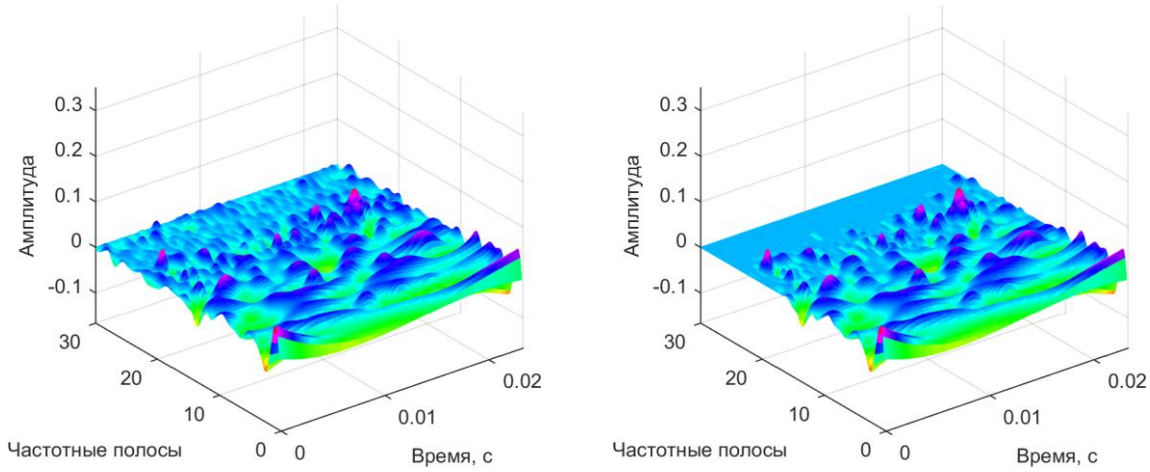
Номер параметра	*X*	*l*	*n*	*k*
1	0,468 38	7	7	4
2	–0,667 70	8	5	3
3	–0,483 42	8	5	1
4	–0,397 76	8	4	1
5	0,284 50	7	5	2

В настоящей работе применяется скалярное квантование атомов. Максимальный шаг квантования определяется согласно выражению (1). После завершения процесса квантования производится энтропийное кодирование на основе алгоритма Хаффмана. Кодовые книги сформированы для каждого уровня предельного дерева декомпозиции ПДВП в силу динамической структуры дерева для каждого входного фрейма аудиосигнала.

Кодирование координат позиции атомов в дереве (*l*, *n*, *k*), что фактически представляет собой кодирование самой структуры дерева текущего фрейма, реализовано следующим образом. Для первого фрейма производится двоичное кодирование всей структуры дерева *E*. Для каждого последующего фрейма кодируется только его отличие от предыдущего. Вариантов, различающих два смежных фрейма, всего два: узел дерева расщепляется либо удаляется. Также встречаются две частые ситуации: узел остается без изменений; несколько раз подряд повторяется ситуация, в которой не происходит изменений. На каждую данную ситуацию выделяется двухразрядный двоичный код, который и передается вместе с закодированными атомами декодеру.

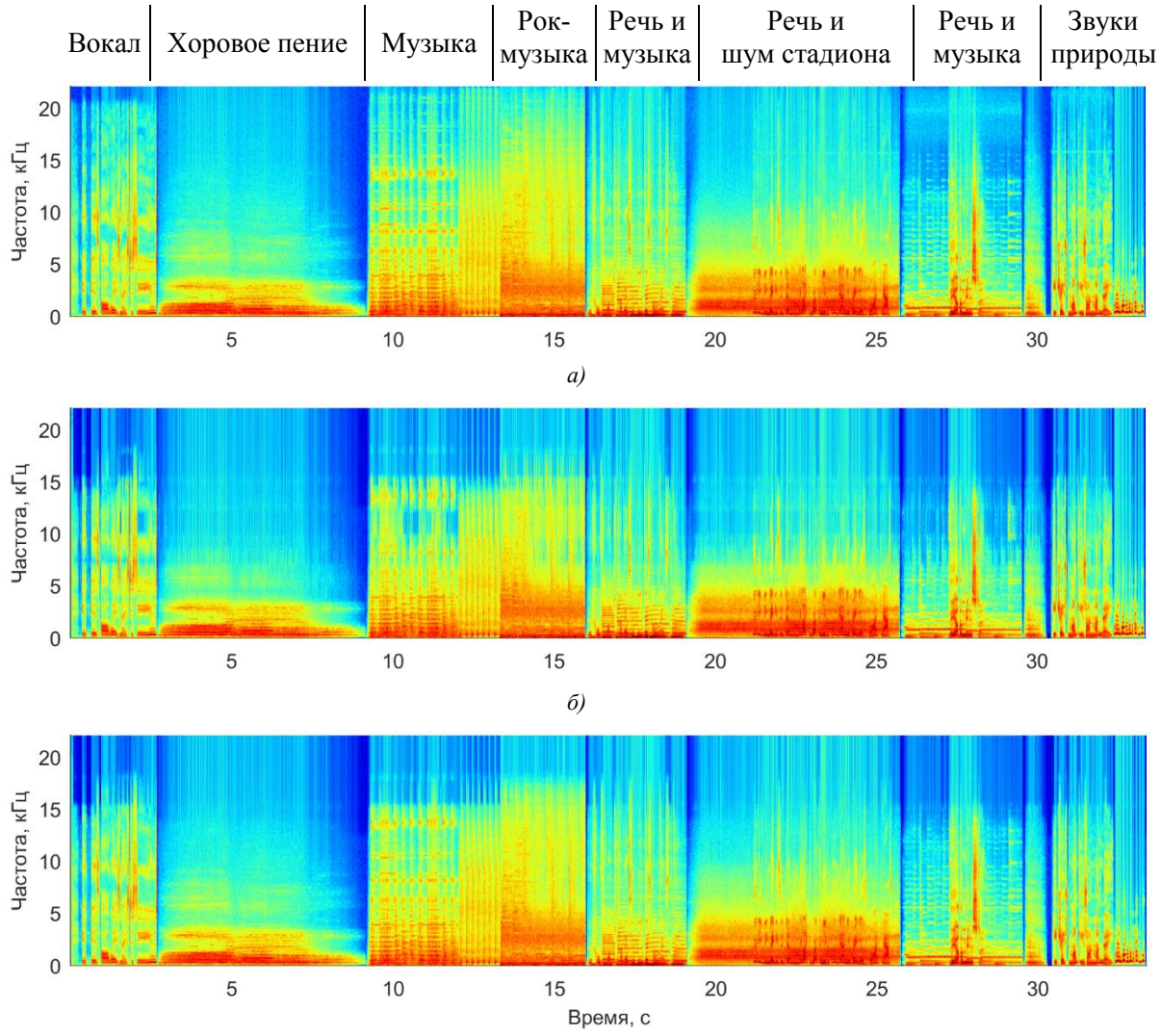
Результаты работы модели кодера и декодера представлены на рис. 9 и 10. На рис. 9 показаны скалограммы аудиторного возбуждения. Скалограмма – это графическое представление коэффициентов вейвлет-преобразования в виде трехмерного графика, где по оси *X* отложена длительность сигнала, по оси *Y* – частотные полосы, а ось *Z* показывает значение коэффициентов анализируемого сигнала. На рис. 9, *б* видно, что алгоритмом СП был произведен отбор максимально значимых для восприятия атомов. Различия между скалограммами оригинального фрейма сигнала и реконструированного отчетливо видны в частотных полосах, соответствующих высокочастотным областям сигнала: наблюдается отсутствие части коэффициентов, которые оказались ниже порога маскирования, а также тех, для которых не хватило бюджета атомов.

На рис. 10 показаны спектрограммы аудиосигналов (оригинального и реконструированного) с различным звуковым информационным наполнением.



а) *б)*

Рис. 9. Скалограммы аудиторного возбуждения: *а)* оригинальный сигнал; *б)* сигнал, реконструированный с помощью 200 атомов



а)

б)

в)

Рис. 10. Результат работы модели кодирования в частотной области: *а)* оригинальный сигнал; *б)* реконструированный с помощью 200 атомов; *в)* реконструированный с помощью 400 атомов

На рис. 10 видно, что при малом количестве атомов, используемых для реконструкции сигнала, некоторая часть спектра не восстанавливается. Как правило, это касается высокочастотных составляющих, потому что их перцептуальный вклад в общую звуковую картину в основном минимален (чаще всего это детализация звука) и при выборе атомов алгоритмом СП коэффициенты, соответствующие данной области, не входят в итоговой набор. Однако с ростом количества используемых для восстановления коэффициентов объем реконструируемой информации увеличивается.

3. Результаты экспериментов и сравнительная оценка с известными алгоритмами сжатия аудиосигналов

Экспериментальные исследования разрабатываемого аудиокодера проводились на 12 образцах (табл. 2) с разнообразным звуковым информационным наполнением. Тестовые аудиосигналы представляют собой одноканальные записи с минимальной длительностью 7 с, частотой дискретизации 44,1 кГц и разрядностью отсчетов 16 битов.

Таблица 2

Тестовые образцы

Тестовый образец	Описание	Тестовый образец	Описание
*es*01	Вокал (Suzan Vega)	*si*01	Клавесин
*es*02	Речь на немецком языке	*si*02	Кастаньеты
*es*03	Речь на английском языке	*si*03	Труба
*sc*01	Соло на трубе и оркестр	*sm*01	Волынка
*sc*02	Оркестровое произведение	*sm*02	Металлофон
*sc*03	Современная поп-музыка	*sm*03	Струнный инструмент

Одной из метрик, позволяющих определить перцептуальное отличие искаженного сигнала от оригинального, является PEAQ (Perceptual Evaluation of Audio Quality) [19]. Данная оценка предполагает использование модели слуховой системы человека (рис. 11).

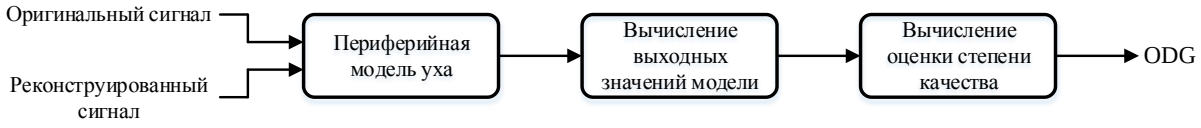


Рис. 11. Общая схема метрики PEAQ

Выходным значением оценки PEAQ является степень объективного различия (ODG, Objective Difference Grade), которая согласно исследованиям имеет высокую степень корреляции с субъективной оценкой качества звука. Шкала ODG в зависимости от степени искажения оцениваемого сигнала формируется следующим образом: 0 – не воспринимаемые на слух искажения; –1 – воспринимаемые, но не раздражающие искажения; –2 – немного раздражающие искажения; –3 – раздражающие; –4 – очень раздражающие искажения сигнала.

В проводимых экспериментах рассчитывалось значение оценки PEAQ – ODG на 12 тестовых образцах из табл. 2. Каждый звуковой сигнал был подвергнут операции кодирования и декодирования в семи вариантах степени сжатия: от 200 атомов, использованных для реконструкции, до 500 атомов с шагом в 50. Расчет скорости битового потока показал, что для 200 атомов ориентировочный битрейт составляет 36,4 кбит/с. Каждые дополнительные 50 атомов (при использовании схемы масштабирования, например) добавляют 8,6 кбит/с к общей скорости битового потока.

На рис. 12 показано распределение значений ODG, сгруппированных для каждого образца по возрастанию битрейта. Из рисунка видно, что ни один образец не получил оценку –4 и ниже. Это значит, что при минимальном использованном в экспериментах количестве атомов

для реконструкции выходного сигнала степень искажения не была «очень раздражающей» (согласно шкале ODG). Только четыре из 12 звуковых сигналов были оценены в области от –3,5 до –4 при использовании 200 атомов: *si*01, *si*03, *sm*01, *sm*02. Более того, при минимальном количестве использованных для реконструкции атомов ODG двух тестовых аудиосигналов находится в области от –1 до –1,5 («воспринимаемые, но не раздражающие искажения»): речь на немецком языке и кастаньеты. Также на рис. 12 видно стабильное увеличение оценки с ростом количества атомов, использованных для реконструкции. Так, уже при 300 атомах половина тестовых образцов находится в области до –1 либо на границе этого значения, а это значит, что оцененная степень искажения фактически является «не воспринимаемой на слух». При 450 атомах только два аудиосигнала оценены в –1,40 и –1,77 (*si*03 и *sm*02 соответственно), остальные оценки уже находятся до границы –1, и при 500 атомах 11 из 12 тестируемых звуковых образцов имеют «не воспринимаемые на слух искажения».

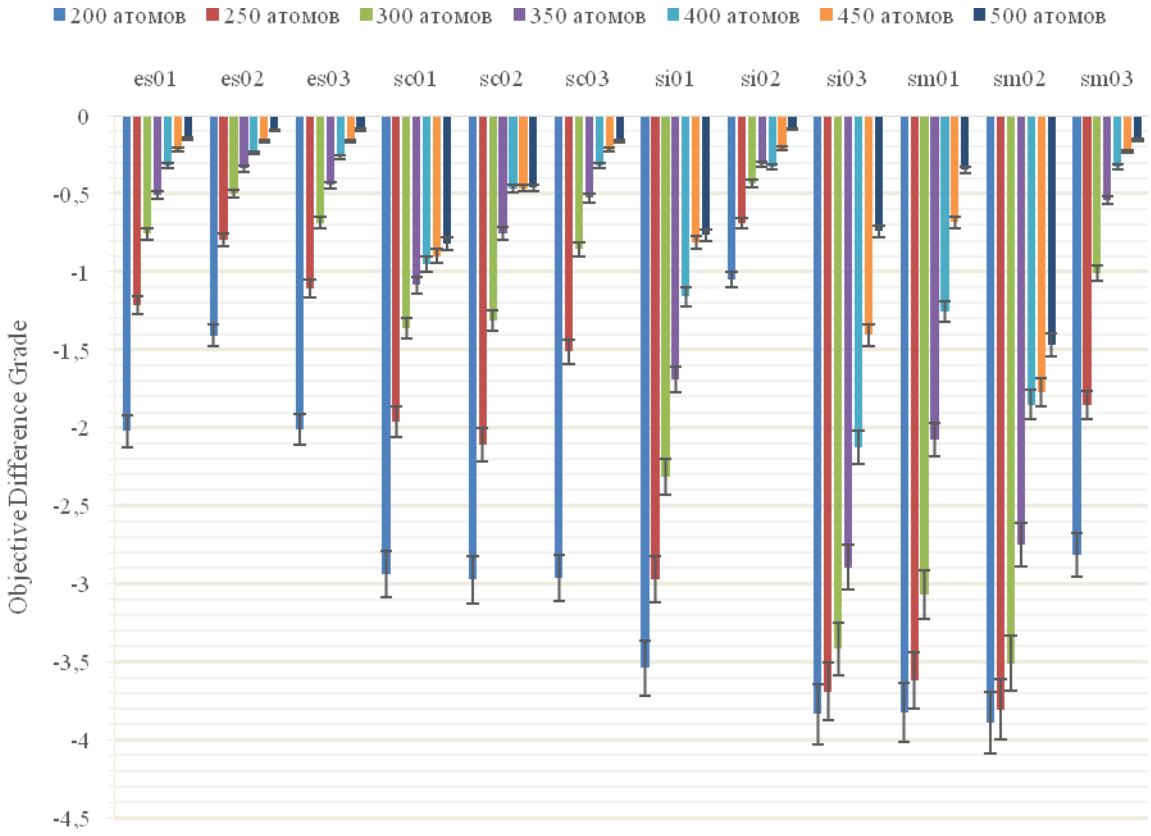


Рис. 12. Результаты объективной оценки качества в соответствии с метрикой PEAQ

Данные результаты говорят о том, что благодаря масштабируемости разрабатываемого аудиокодера есть возможность настраивать скорость битового потока (т. е. степень сжатия сигнала) под конкретную ситуацию без потери качества восстановленного выходного сигнала. Например, в зависимости от информационного наполнения входных данных можно отрегулировать битрейт: как видно на рис. 12, все речевые сигналы (*es*02, *es*03) уже при 250 атомах (расчетный битрейт 45 кбит/с) обладают оценками до –1,1, т. е. «не воспринимаемые на слух искажения» и немного выше данной границы.

На рис. 13 видно, что при минимальном битрейте средняя оценка предлагаемого кодера и кодеров Opus и Vorbis [20, 21] находится в диапазоне от –3 до –2 («немного раздражающие искажения»). С ростом скорости битового потока наблюдается стабильное увеличение качества реконструированного аудиосигнала. Начиная с области 71 кбит/с, все кодеры, включая разрабатываемый, показывают приблизительно одинаковые значения оценки ODG, находясь в области «не воспринимаемые на слух искажения». Все три сравниваемых аудиокодера имеют прибли-

зительно эквивалентные оценки ODG, однако стоит отметить, что Vorbis является архивным кодером, в то время как разрабатываемый кодер обладает возможностью функционировать в реальном масштабе времени. Opus представляет собой универсальный кодер реального времени, но в его составе находится детектор входного сигнала и две различные модели для работы с речевыми и другими звуковыми данными, что может осложнить его реализацию на целевой платформе. По сравнению с ним разрабатываемый аудиокодер содержит в своем составе одну модель для работы со всеми входными аудиоданными.

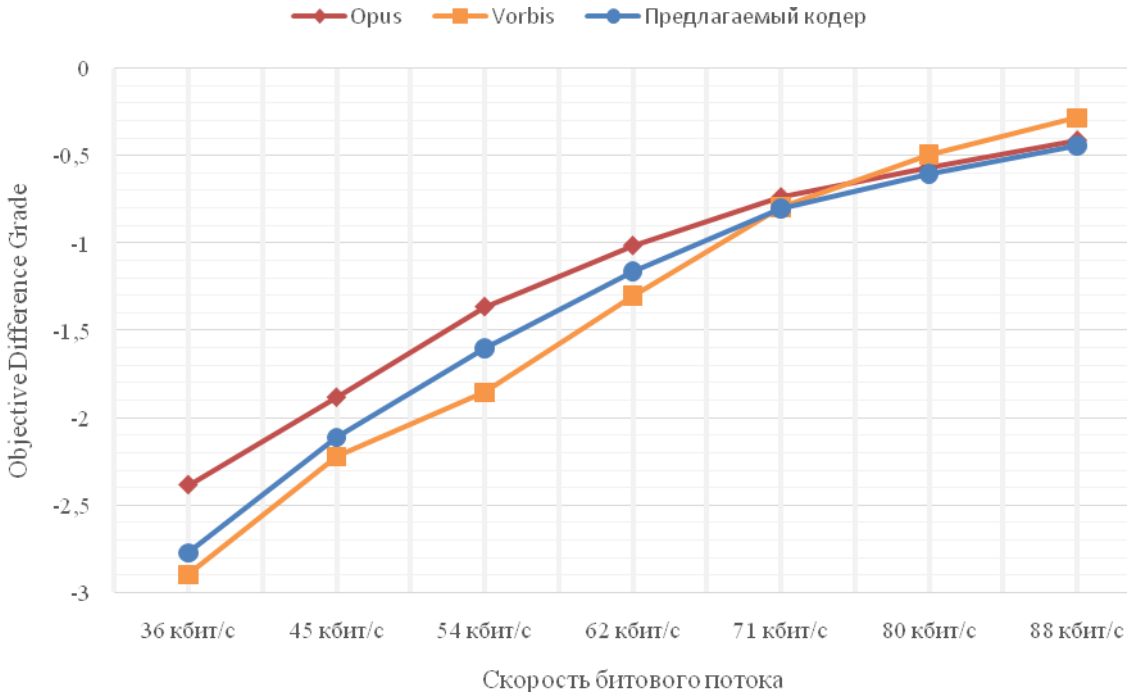


Рис. 13. Сравнение средней оценки разрабатываемого аудиокодера с кодерами Opus и Vorbis

Заключение

В работе предложен алгоритм универсального масштабируемого аудиокодера на основе разреженной аппроксимации с динамически оптимизируемым от фрейма к фрейму согласно психоакустическому критерию словарем частотно-временных функций на базе ПДВП. Описаны принципы адаптации частотно-временного плана к входному фрейму сигнала на основе двух стоимостных функций: перцептуальной и временной энтропии. Дано пояснение работы процедуры СП с целью выбора наиболее важных для восприятия слуховой системой человека коэффициентов входного сигнала. Описана структура работы алгоритмов кодирования и декодирования сигнала на базе данной модели. Проведенный сравнительный анализ показал, что разрабатываемый аудиокодер способен эффективно (с высоким качеством восстановленного сигнала при низких скоростях битового потока) работать со звуковыми сигналами различного информационного наполнения в реальном масштабе времени. Кроме того, возможность масштабирования позволяет адаптировать передаваемое либо сохраняемое количество информации в зависимости от заданного ресурса или природы обрабатываемого входного сигнала. Сравнение с современными звуковыми кодерами Opus и Vorbis показало, что представленный алгоритм сжатия обладает эквивалентным качеством восстановленного сигнала для большинства тестовых образцов, а за счет возможности масштабирования при небольшом увеличении скорости битового потока позволяет добиться сравнимых показателей и для остальных аудиосигналов.

Направление дальнейших исследований будет следующим: изучение возможности оптимизации процедуры СП для повышения быстродействия алгоритма, а также уточнения выбира-

емых из сигнала коэффициентов (атомов) для увеличения качества выходного сигнала и снижения при этом скорости битового потока; реализация возможности работы с переменной длиной окна анализа (длительности фрейма), что может привести к более точной локализации транзиентных компонентов сигнала; разработка эффективной схемы квантования отобранных атомов, что позволит увеличить степень компрессии выходного аудиосигнала.

Список литературы

1. Mallat, S. Matching pursuit with time-frequency dictionaries / S. Mallat, Z. Zhang // IEEE Transactions on Signal Processing. – December, 1993. – Vol. 41, no. 12. – P. 3397–3415.

2. Petrovsky, Al. Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding / Al. Petrovsky, E. Azarov, A. Petrovsky // Elsevier, Signal Processing. Special «Issue Fourier Related Transforms for Non-Stationary Signals». – June 2011. – Vol. 91, iss. 6. – P. 1489–1504.

3. Ruiz Reyes, N. Adaptive signal modelling based on sparse approximations for scalable parametric audio coding / N. Ruiz-Reyes, P. Vera Candeas // IEEE Transactions on audio, speech and language processing. – 2010. – Vol. 18, iss. 3. – P. 447–460.

4. Chardon, G. Perceptual matching pursuit with Gabor dictionaries and Time-Frequency Masking / G. Chardon, T. Necciari, P. Balazs // ICASSP'2014. – Florence, Italy, 2014. – P. 3126–3130.

5. Ravelli, E. Union of MDCT bases for audio coding / E. Ravelli, G. Richard, L. Daudet // IEEE Transactions on audio, speech and language processing. – 2008. – Vol. 16, iss. 8. – P. 1361–1372.

6. Mallat, S.A. Wavelet Tour of Signal Processing. The Sparse Way; 3rd ed. / S.A. Mallat. – Burlington, MA : Academic Press, 2008. – 832 p.

7. Strang, H. Wavelets and Filter Banks / H. Strang, T. Nguyen. – Wellesley, MA : Wellesley-Cambridge Press, 1997. – 520 p.

8. Petrovsky, Al. Scalable parametric audio coder using sparse approximation with frame-to-frame perceptually optimized wavelet packet based dictionary / Al. Petrovsky, V. Herasimovich, A. Petrovsky // AES 138th Convention. – Warsaw, Poland, 2015. – Paper 9264.

9. Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика (с MATLAB-примерами) / под ред. А.А. Петровского. – Минск : Бестпринт, 2009. – 456 с.

10. Daubechies, I. Ten lectures on Wavelets / I. Daubechies. – Philadelphia, Pennsylvania : Society for industrial and applied mathematics, 1992. – 357 p.

11. Johnston, J.D. Transform coding of audio signals using perceptual noise criteria / J.D. Johnston // IEEE Journal on Selected Areas in Communications. – February 1988. – Vol. 6, iss. 2. – P. 314–323.

12. Петровский, Ал.А. Построение психоакустической модели в области вейвлет-коэффициентов для перцептуальной обработки звуковых и речевых сигналов / Ал.А. Петровский // Речевые технологии. – 2008. – № 4. – С. 61–71.

13. Painter, T. Perceptual Coding of Digital Audio / T. Painter, A. Spanias // Proceedings of the IEEE. – April 2000. – Vol. 88, iss. 4. – P. 451–515.

14. Umapathy, K. Audio signal processing using time-frequency approaches: coding, classification, fingerprinting, and watermarking / K. Umapathy, B. Ghoraani, S. Krishnan // EURASIP Journal on Advances in Signal Processing. – 2010. – Vol. 2010. – P. 1–28.

15. Goodwin, M. Atomic decompositions of audio signals / M. Goodwin, M. Vetterli // Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics. – New Paltz, NY, USA, 1997. – P. 1–4.

16. Петровский, Ал.А. Масштабируемые аудиоречевые кодеры на основе адаптивного частотно-временного анализа звуковых сигналов / Ал.А. Петровский, А.А. Петровский // Труды СПИИРАН. – 2017. – № 1(50). – С. 55–92.

17. Petrovsky, Al. Audio/speech coding using the matching pursuit with frame-based psychoacoustic optimized time-frequency dictionaries and its performance evaluation / Al. Petrovsky, V. Herasimovich, A. Petrovsky // Signal Processing: Algorithms, Architectures, Arrangement, and Applications (SPA). – Poznan, Poland, 2016. – P. 225–229.

18. Petrovsky, A. Real-time wavelet packet-based low bit rate audio coding on a dynamic reconfiguration system / A. Petrovsky, D. Krahe, A.A. Petrovsky // AES 114th Convention. – Amsterdam, 2003. – Paper 5778.

19. ITU-R Rec. BS.1387-1, Method for objective measurements of perceived audio quality, 2001.

20. High-quality, low-delay music coding in the Opus codec / J.-M. Valin [et al.] // AES 135th Convention. – NY, USA, 2013. – Paper 8942.

21. Voice coding with Opus / K. Vos [et al.] // AES 135th Convention. – NY, USA, 2013. – Paper 8941.

Поступила 18.10.2017

Белорусский государственный университет
информатики и радиоэлектроники,
Минск, П. Бровки, 6
e-mail: gerasimovich@bsuir.by,
alexey@petrovsky.eu

V.Y. Herasimovich, Al.A. Petrovsky

PSYCHOACOUSTICALLY MOTIVATED TIME-FREQUENCY DICTIONARY BUILDING FOR UNIVERSAL SCALABLE AUDIOCODER BASED ON THE SPARSE APPROXIMATION

The article studies the process of creating a perceptually-motivated dictionary of the time-frequency functions based on the wavelet packet transform optimized for the input signal frame and its utilization in the universal scalable real-time audiocoder. The article points out the importance of the topic, great attention is paid to the psychoacoustic modelling. It describes the following algorithms: sparse approximation, perceptual adaptation of the wavelet packet decomposition tree, input signal encoding/decoding schemes. The results of the experimental research of the developed coding algorithm and comparison with the modern coding schemes such as Opus and Vorbis based on the objective quality assessment PEAQ – ODG were also given.