УДК 004.032.6 **DOI** 10.15622/sp.50.3

Ал.А. Петровский, А.А. Петровский

МАСШТАБИРУЕМЫЕ АУДИОРЕЧЕВЫЕ КОДЕРЫ НА ОСНОВЕ АДАПТИВНОГО ЧАСТОТНО-ВРЕМЕННОГО АНАЛИЗА ЗВУКОВЫХ СИГНАЛОВ

Петровский Ал.А., Петровский А.А. **Масштабируемые аудиоречевые кодеры на основе адаптивного частотно-временного анализа звуковых сигналов.**

Аннотация. Рассматриваются методы перцептуальной субполосной обработки звуковых сигналов с динамической трансформацией частотно-временного плана на основе пакетного дискретного вейвлет-преобразования (ПДВП), достоинством которых является то, что рост дерева осуществляется сверху вниз, без возвратов на меньшие масштабные уровни преобразования и необходимости построения полного дерева ПДВП, что соответствует концепции реализации масштабируемых аудиоречевых кодеров в реальном масштабе времени. Приводятся объективные оценки качества предлагаемых кодеров на основе методики *PEMO-Q* и сравнения с широко распространенными кодерами *Opus* и *Vorbis*, которые показывают, что реконструированный сигнал соответствует требованиям стандарта ITU-R PEAQ при высокой степени компрессии в 18 и более раз, не содержит артефактов: отношение мощности шума к порогу маскирования NMR_{total} меньше ≈ –9 дБ.

Ключевые слова: масштабируемый аудиоречевой кодер, пакет дискретного вейвлет-преобразования, согласованная подгонка.

1. Введение. Термин «мультимедиа» в настоящее время приобрел необычайную популярность. Это объясняется тем, что появилась возможность связывать в единое целое визуальную, акустическую и текстовую информацию, то есть сделан большой шаг от простых массовых продуктов, например, таких как сотовый телефон и MP3-проигрыватель, к амбициозным проектам — человеко-машинным интерфейсам. Стремительное внедрение цифровых технологий в такие приложения, как аудиокниги, системы радиовещания и телевидения, непрерывное увеличение передач через Интернет, информационное наполнение которых может быть различно и не ограничено только речью или музыкой, определило необходимость в объединенном аудиоречевом кодере с низкой скоростью цифрового потока, который одинаково хорошо работает со всеми типами звукового информационного наполнения [1-3].

В большинстве своем техника обработки аудиосигналов связана с тремя областями: компрессией, классификацией и защитой информации. Компрессия речевых сигналов традиционно базируется на определенных моделях речеобразования, в то время как в методах высококачественного кодирования аудиосигналов используется свойство шумового маскирования человеческого слуха [4-6]. Общая философия перцептуального кодера взаимосвязана с выбором метода частотно-временного анализа — банка цифровых фильтров [7]. Развитие этих работ в настоящее время

идет по пути построения перцептуальных субполосных аудиокодеров с постоянным частотно-временным планом, что приводит к высокой скорости цифрового потока для достижения высокого качества реконструированного сигнала [8]. Увеличение степени компрессии обусловливает применение техник частотно-временного анализа с динамически изменяемым частотно-временным планом для соответствующего фрейма обработки аудиосигнала, например, наиболее гибкого в смысле управления частотно-временным разрешением, пакета дискретного вейвлет-преобразования (ПДВП) [9]. Адаптивная частотно-временная аппроксимация аудиосигнала на основе ПДВП с перцептуально-оптимизированным частотно-временным планом позволяет исключить избыточность в сигнале, что обусловлено невосприимчивостью определенных частотных компонент человеком, вследствие маскирующего эффекта [10, 11].

В настоящее время для решения задачи компрессии звуковых сигналов часто применяются алгоритмы разреженной аппроксимации на основе согласованной подгонки (СП) [12, 13], заключающейся в поиске наилучшей проекции входного сигнала на избыточный словарь базисных функций — атомов [14]. В работе [15] представлен метод разреженной аппроксимации на основе СП со словарем, состоящим из функций Габора. Данный алгоритм позволяет достичь почти перфективной реконструкции входного сигнала, но при большом числе итераций. Кроме того, словарь атомов имеет фиксированную структуру, что влечет за собой увеличение его размера и вычислительной сложности алгоритма. В подходе [16] используется два словаря, переключение между которыми зависит от энергии остаточного сигнала. Это дает большую гибкость алгоритму и уменьшает его вычислительную сложность, однако параметр переключения между словарями фиксированный, что не всегда адекватно обрабатываемому фрейму входного сигнала. Гибридная модель представления аудиосигнала на основе трех частей: гармонической, шумовой и переходной применяется в масштабируемых аудиокодерах [12, 17]. Переходная составляющая параметризируется с помощью СП. Данный вариант кодирования требует трех различных алгоритмов для параметризации сигнала, что обуславливает большую вычислительную сложность метода. В [13, 18, 19] решается задача параметрического анализа аудиосигналов на основе разреженной аппроксимации с перцептуально-оптимизированным ПДВП-словарем вейвлет-коэффициентов, применение которого, как показано в [19], дает возможность построения универсального аудиоречевого кодера.

Цель данной работы — показать методы перцептуальной субполосной обработки звуковых сигналов на основе ПДВП с динамической трансформацией частотно-временного плана анализа для построения масштабируемых аудиоречевых кодеров.

2. Частотно-временной анализ с использованием пакетного дискретного вейвлет-преобразования.

2.1. Пакетное дискретное вейвлет-преобразование. Пусть $\{\psi_n(t): n \in Z\}$ определяет множество структур деревьев ПДВП и пусть $E \in \{(l,n)\}: 0 \le l \le L$ — масштабный уровень, $0 \le n \le 2^l$ — номер узла на масштабном уровне представляет собой узлы дерева ПДВП, тогда отрезок $[0,1)$ разделяется на диадические интервалы $I_{l,n} = [n2^{-l}, (n+1)2^{-l}]$, которые соответствуют специфическому множеству узлов E. В частности, $\{\psi_{l,n,k}(t): (l,n) \in E, k \in Z\}$, где $\psi_{l,n,k}(t) \triangleq 2^{-l/2}\psi_n(2^{-1}t-k)$ является базовой формой в пространстве сигнала $\overline{span}\{\psi_0(t-k): k \in Z\}$. Узел $(l,n) \in E$ дерева ПДВП ассоциируется с частотной полосой, у которой центральная частота и полоса пропускания приблизительно задаются следующими соотношениями: $f_{l,n} = 2^{-l}(GC^{-1}(n)+0.5)\cdot f_s/2, \Delta f_{l,n} = 2^{-l}\cdot f_s/2$, где GC^{-1} — обратный код перестановок Грея, f_s — частота дискретизации сигнала. В отличие от дискретного вейвлет-преобразования, представляющего собой дерево декомпозиции области нижних частот, ПДВП это декомпозиция обеих частей дерева преобразования: области нижних и верхних частот [20].

На рисунке 1 (а, в, д, ж) показаны возможные варианты построения второй ступени ПДВП. Как видно из рисунка 1, структура дерева преобразования позволяет варьировать разрешающую способность в частотной и временной областях. Выполнение более детальной декомпозиции, например, в области нижних частот, приводит к увеличению разрешающей способности в области частот и ее уменьшению во временной области, в то время как менее детальная декомпозиция в области верхних частот, обеспечивает высокое разрешение во временной области и уменьшает частотное разрешение. Соответствующие частотно-временные планы показаны на рисунке 1 (б, г, е, з), где кружками отмечены коэффициенты преобразования $X_{l,n,k}(t)$. Здесь индексы имеют следующее назначение: l — уровень преобразования, n — номер узла на уровне l; k — это временная позиция в блоке обработки. Анализ входного сигнала $x(t)$ на базе ПДВП с переменным частотно-временным разрешением (многомасштабный анализ [20]) выполняется на основе рекурсивных масштабной и вейвлет-функций. При цифровой реализации это коэффициенты цифровых фильтров с конечной импульсной характеристикой.

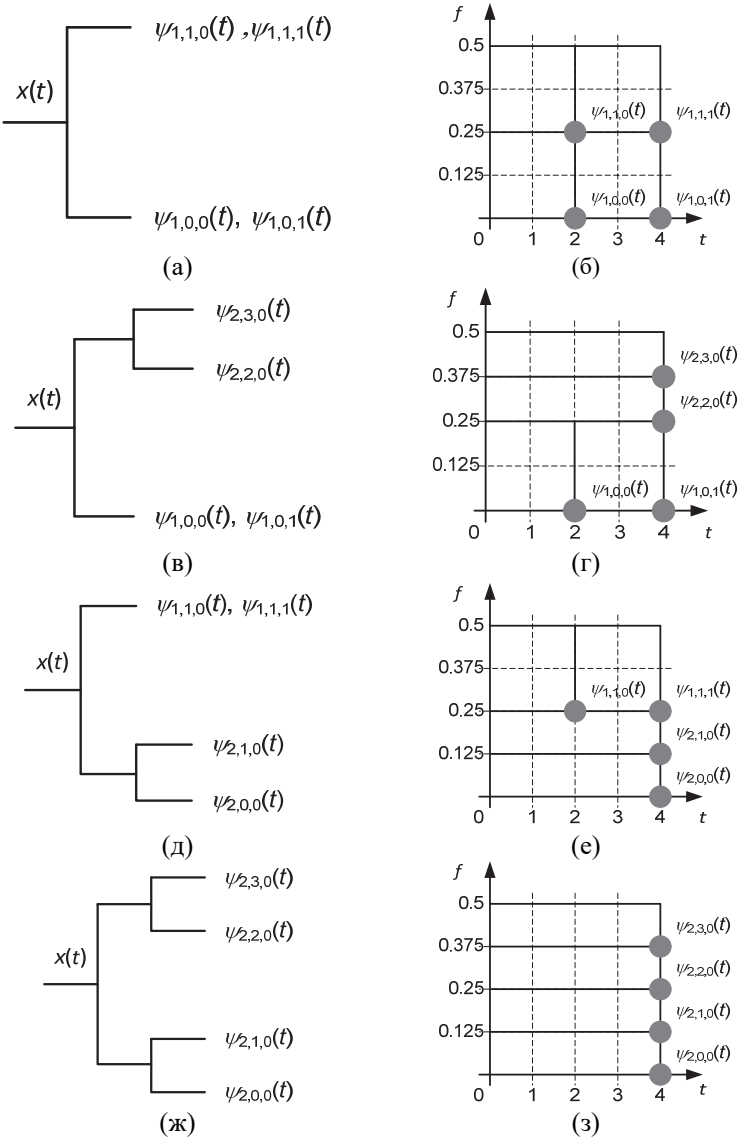


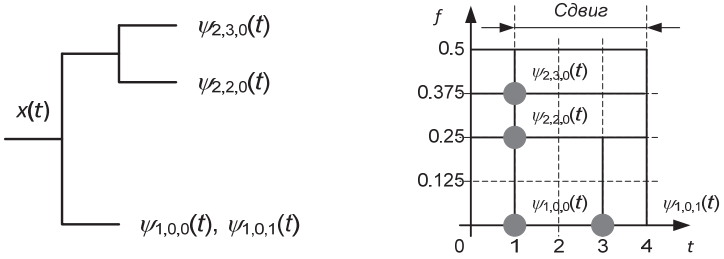
Рис. 1. Структуры деревьев ПДВП: (а), (в), (д), (ж) – возможные варианты построения второй ступени ПДВП; (б), (г), (е), (з) – частотно-временные планы соответствующие вариантам ПДВП

Таким образом, в отличие от дискретного преобразования Фурье (ДПФ) и банков ДПФ или косинусно модулированных фильтров [7] ПДВП обеспечивает двумерную развертку кодируемого звукового сигнала, при этом частота и время рассматриваются как независимые переменные. В результате появляется возможность анализировать свойства сигнала одновременно в частотном и временном пространствах.

2.2. ПДВП с управляемым временным сдвигом. Декомпозиция ПДВП может быть расширена в корень из двух раз путем варьирования временным сдвигом в каждом из узлов n структуры дерева ПДВП $l, n \in E$. Увеличение числа возможных комбинаций не изменяет древовидную структуру ПДВП. Временной сдвиг модифицирует ортогональные проекции ортонормальных масштабных функций на подпространстве $V_{l,n}$ [7, 21]. Данное изменение описывает как уже существующие, так и вновь образованные подмножества:

$$V_{l,n,m} \equiv \left\{ \psi_{l,2n+1,m}(t) = 2^{l/2} \sum_{k \in Z} \psi_{l,n,m}\left(2\left(t - m_c\right) - k\right) \right\}, \qquad (1)$$

где $m_c = \{m, m + 2^{-l}\}$ — временной сдвиг, $l, n \in E$. Тройные индексы каждого узла (l, n, m) структуры дерева E ПДВП определяют номер узла n, уровень l, временной сдвиг m. Декомпозиция сигнала на основе ПДВП без временного сдвига, например, (рисунок 1в и 1г), а соответствующая ей структура и частотно-временной план ПДВП для выполнения декомпозиции с временным сдвигом показаны на рисунке 2, где структура ПДВП получается путем каскадного соединения блоков декомпозиций удовлетворяющих записи в выражении (1).



(а) (б)

Рис. 2. Декомпозиция ПДВП с временным сдвигом: а) вариант построения второй ступени ПДВП; (б) частотно-временной план соответствующий варианту ПДВП

Как видно из рисунка 2, структура дерева ПДВП позволяет не только варьировать разрешающую способность в частотной и временной областях, но и выполнять сдвиг по времени для более детальной декомпозиции входного сигнала [21, 22]. На рисунках 3 и 4 показаны структуры деревьев ПДВП и соответствующие частотно-временные планы для тестового сигнала, содержащего короткие импульсы тона с нелинейной частотной модуляцией для адаптивного ПДВП без временного сдвига и с временным сдвигом. Как видно из данного примера, адаптивный ПДВП с временным сдвигом лучше локализовал неопределенность (информативность) сигнала, чем без сдвига.

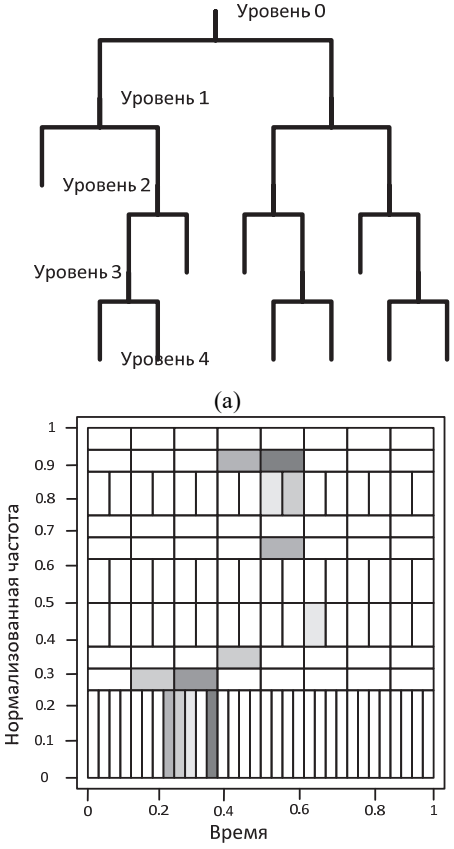


Рис. 3. Декомпозиция сигнала на основе ПДВП без временного сдвига: (а) структура ПДВП; (б) частотно-временной план ПДВП

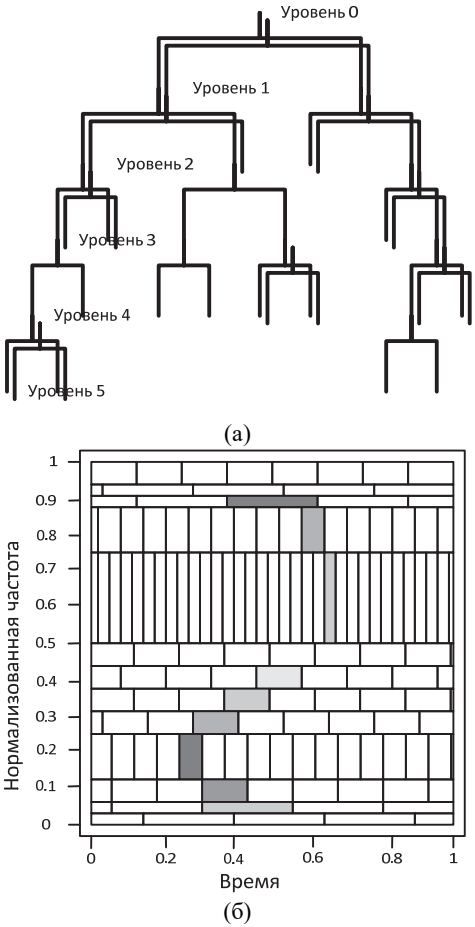


Рис. 4. Декомпозиция сигнала на основе ПДВП с временным сдвигом: (а) структура ПДВП; (б) частотно-временной план ПДВП

2.3. ПДВП согласованный с критической частотной шкалой. Построение частотно-временного плана для перцептуальной обработки сигналов речи и звука, согласованного с критической частотной шкалой, при минимальном множестве узлов ПДВП, определенной глубине декомпозиции структуры дерева, заданной частоте дискретизации сигнала, минимальной ошибке аппроксимации критических частотных полос в области барков [11] осуществляется в соответствии с психоакустической моделью Zwicker [23], где:

— расстояние между центральными частотами соседних критических частотных полос в барках:

$$Z = F(f) = 13 \cdot \arctan(0.00076 \cdot f) + 3.5 \cdot \arctan\left(\left(\frac{f}{7500}\right)^2\right), \qquad (2)$$

где f — частота в герцах, единица измерения в данном масштабе — 1 барк;

— ширина критических частотных полос в герцах.

$$CBW(f) = 25 + 75 \cdot (1 + 1.4 \cdot (f / 1000)^2)^{0.69}. \qquad (3)$$

При этом, также вычисляются характеристики психоакустической модели восприятия человеком акустической информации, такие как пороги маскирования:

— абсолютный порог слышимости ATH (absolute threshold of hearing), частотная зависимость которого аппроксимируется выражением:

$$ATH_{SPL}(f) = 3.64 \cdot (10^{-3} \cdot f)^{-0.8} - 6.5 \cdot \exp(-0.6 \cdot (10^{-3} \cdot f - 33)^2) + 10^{-3} \cdot (10^{-3} \cdot f)^4, [\text{дБ}] \qquad (4)$$

где f — частота в герцах;

— частотное маскирование (simultaneous masking), проявляющееся при воздействии маскера в течение некоторого времени одновременно на разных частотах сигнала;

— маскирование во временной области (temporal masking): если громкий звук маскирует следующий за ним слабый звук, то такое явление называется маскированием вперед (post-masking), которое может продолжаться от 5 до 300 мс. в зависимости от силы и длительности звука; маскирование назад (pre-masking), когда громкий звук маскирует звук, воспроизводимый до него, длительность которого составляет примерно до 20 мс.

Для того чтобы получить аппроксимацию шкалы критических частотных полос с помощью ПДВП, необходимо осуществить декомпозицию дерева ПДВП таким образом, чтобы расстояние между центрами одной субполосы и другой субполосы составляло 1 барк. Следует отметить, что ширина критических частотных полос $CBW(f)$ монотонно увеличивающаяся функция частоты (3). Для формирования низкоча-

стотных полос требуется интенсивная декомпозиция ПДВП в сравнении с характером изменения дерева ПДВП для аппроксимации высокочастотных полос.

Интегральная перцептуально взвешенная ошибка аппроксимации шкалы критических частотных полос деревом $(l,n) \in E_m$ ПДВП в области Барков может быть определена следующим образом:

$$Q_E = \frac{1}{L} \sum_{\substack{для \\ \forall (l,n) \in E_m}} \left[\widehat{CBW}_{Zw}(Z) - \widehat{CBW}_{E_m}\left(Z_{(l,n)}\right) \right]^2 \cdot \hat{W}(Z). \quad (5)$$

Здесь ширина критических частотных полос $\widehat{CBW}(Z)$ в Гц как функция центральных частот соседних критических частотных полос заданных в барках, то есть:

$$\widehat{CBW}(Z) = CBW\left(F^{-1}(Z)\right), [Гц]. \quad (6)$$

$(\widehat{CBW}_{Zw} Z)$ определяет шкалу критических частотных полос в модели Zwicker [23], $\widehat{CBW}_{E_m}(Z_{(l,n)})$ — аппроксимация критических частотных полос деревом ПДВП $(l,n) \in E_m$, центр $Z_{(l,n)}$ в барках полосы (l,n) дерева ПДВП E_m вычисляется для центральной частоты $f_{(l,n)}$, заданной в Гц, как $Z_{(l,n)\in E_m} = F(f_{(l,n)})$, где F — преобразование (2). Перцептуальная взвешивающая функция $\widehat{W}(Z)$, учитывающая определенные частотные свойства наружного и среднего уха, задает распределение ошибки аппроксимации шкалы критических частотных полос меньше в области средних частот по сравнению с низкочастотным и высокочастотным диапазонами, и определяется в шкале дБ, как функция частоты [24]:

$$W_{дБ}(f) = -0.6 \cdot 3.64 \cdot (10^{-3} \cdot f)^{-0.8} + 6.5 \cdot \exp(-0.6 \cdot (10^{-3} \cdot f - 3.3)^2) - 10^{-3}(10^{-3} \cdot f)^4, \quad (7)$$

а также $\widehat{W}(Z)$ может быть переопределена для шкалы барков, как:

$$\widehat{W}(Z) = \widehat{W}\left(F(f)\right) = W\left(F^{-1}(Z)\right) = W(f), \quad (8)$$

где $W(f) = 10\left(\frac{w_{дБ}(f)}{20}\right)$. Минимизация ошибки Q_E может позволить автоматизировать процесс построения оптимального дерева

ПДВП $(l,\mathrm{n}) \in E_{CB}$ для шкалы критических частотных полос. На рисунке 5 показано дерево ПДВП (Critical Band Wavelet Packet Decomposition $(CB-WPD)$), полученное эмпирически [22, 25, 26].

Дерево $CB-WPD$ делит частотный диапазон [0 – 22,05 кГц] на 25 неравномерных полос $CBW(f)$ [26], то есть на 25 барков. Корневой узел $(l,n)=(0,0)$ данного дерева соответствует всему частотному диапазону сигнала. Каждый внутренний узел дерева $(l,n) \in E$, названный узлом предка, делится на два потомка: 1-й потомок и 2-й потомок, ассоциируемые с высокочастотной и низкочастотной фильтрацией, выходные сигналы (вейвлет-коэффициенты) которых децимируются в соотношении 2:1:

$$X_{l,n,k}(t) = \left\langle x(t), \varphi_{l,n,k}(t) \right\rangle, (l,n) \in E_{CB}, k \in \mathbb{Z}, \qquad (9)$$

где l — номер масштабного уровня преобразования $(0 \le l \le 8)$, n — номер узла масштабного уровня преобразования, k — вейвлет-коэффициентов в полосе (узле (l,n) дерева E). Банк вейвлет-фильтров $(CB-WPD:(l,n) \in E_{CB})$, согласованный с критической шкалой частот восприятия акустической информации человеком, является предельной структурой для метода перцептуального кодирования сигнала звука [26-28].

На рисунках 6 и 7 показаны аппроксимации центральной частоты и ширины каждой частотной полосы критической шкалы частот деревом ПДВП, структура которого приведена на рисунке 5.

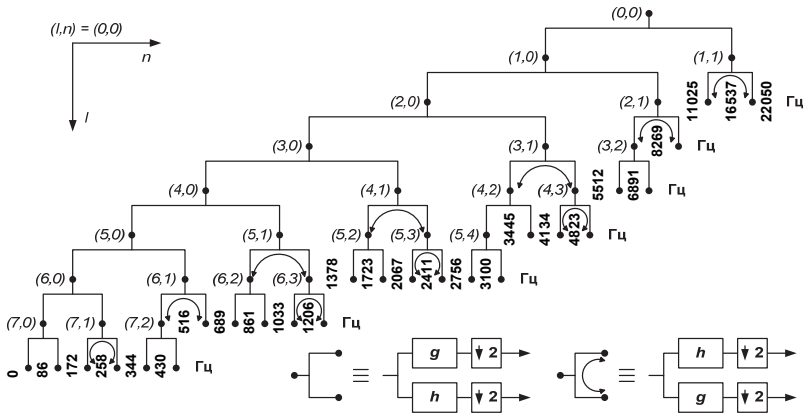


Рис. 5. Структура критического дерева ПДВП $CB-WPD: l=\overline{0,8}$

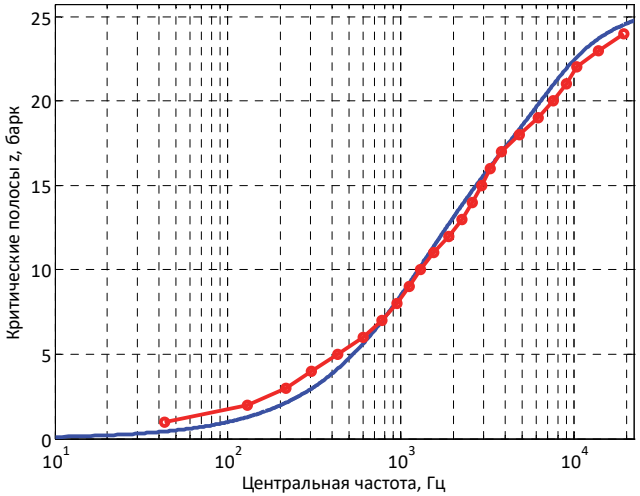


Рис. 6. Аппроксимация центральных частот $CB–WPD:\overline{l=0,8}$

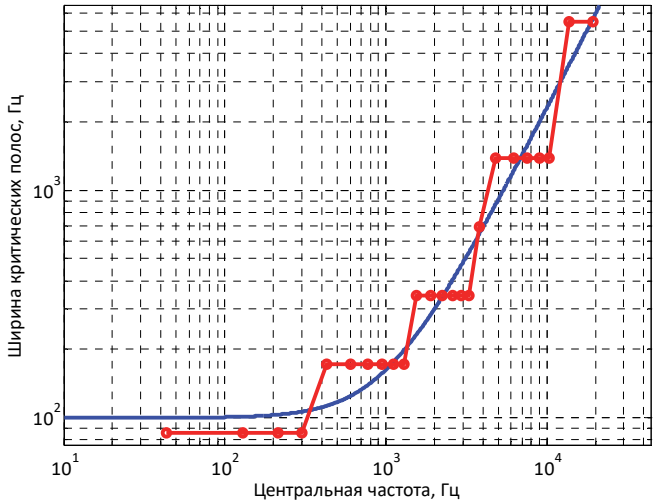


Рис. 7. Аппроксимация ширины критических частотных полос $CB–WPD:\overline{l=0,8}$

Частотно-временной план для структуры дерева ПДВП $CB-WPD:(l,n)\in E_{CB}, l=\overline{0.8}, f_s=44.1\kappa Гц$ (рисунок 5) [26, 29] представлен на рисунке 8.

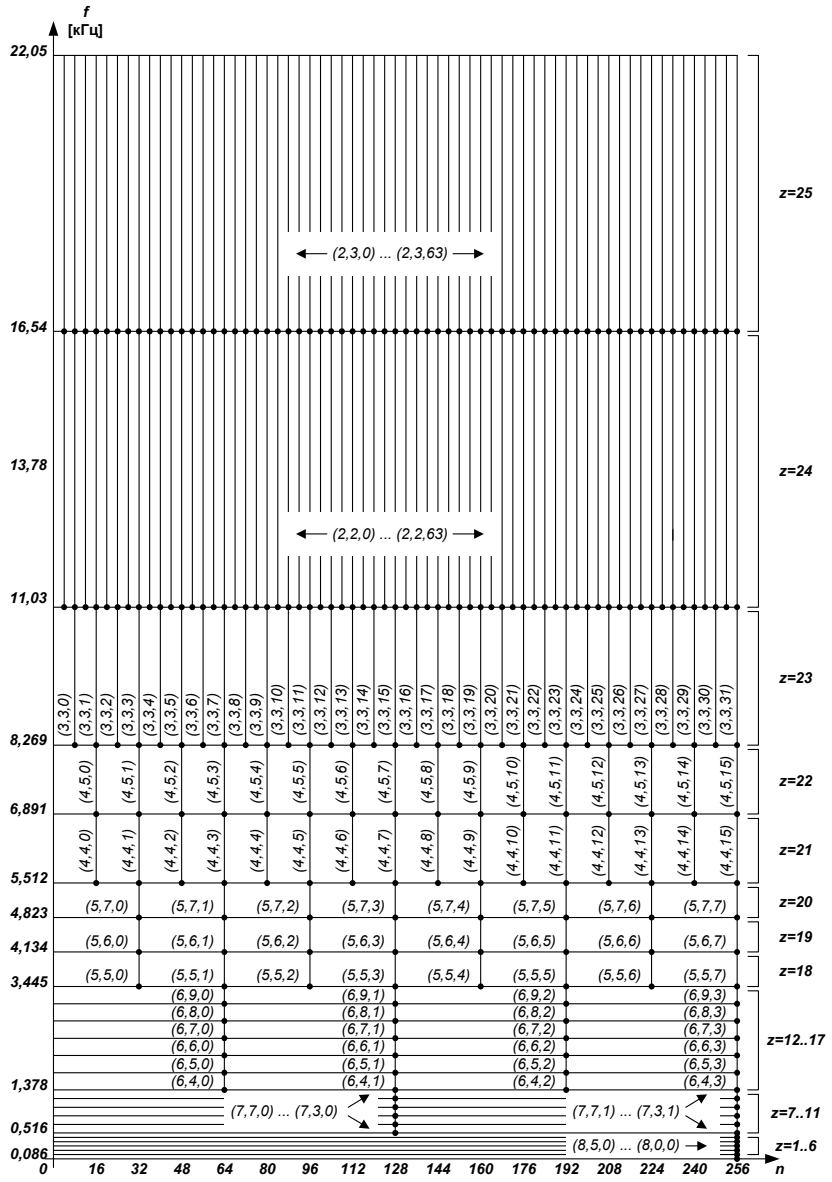


Рис.8. Частотно-временной план структуры дерева ПДВП

$$CB - WPD : l = \overline{0.8}, f_s = 44.1 кГц$$

Ширина каждой клеточки есть длина фрейма, и определяется как $F_l = 2^l \cdot (F_{min} = 2$ отсчетам и $F_{max} = 256$ отсчетам). Следовательно, длина анализируемого окна равна $W = (P-1)(F_{l-1})+1$ отсчетов. Для первого уровня $l = 1$ преобразования определяющей является область верхних частот и длина окна $W = 40$ отсчетам при длине фильтра прототипа $P = 40$. Для уровня $l = 8$ преобразования наибольшая частотная разрешающая способность в области нижних частот, а окно $W = 9946$ отсчетам.

3. Динамическая трансформация частотно-временного плана. В перцептуальном кодировании аудиосигналов выбирается такая декомпозиция ПДВП, при которой минимизируется скорость передачи с сохранением высокого качества восприятия человеком декодированного сигнала. Декомпозиция «лучшего» дерева преобразования выбирается как можно ближе к шкале барков, то есть к шкале критических частотных полос, а банк вейвлет-фильтров, согласованный с критической шкалой частот восприятия акустической информации человеком, является предельной структурой для метода перцептуального кодирования аудио сигнала.

На основе введенной функции стоимости можно определить наилучший путь по этому дереву [30]. Если исходный блок вейвлет-фильтров был ортогональным, то и схема, соответствующая любой конфигурации, будет ортогональной, так как она есть не что иное, как каскадное соединение ортогональных блоков, то есть получается базис, адаптированный к сигналу. В работе [26] предложен метод роста дерева ПДВП или, другими словами, динамическая трансформация частотно-временного плана, позволяющий определить субоптимальную структуру декомпозиции ПДВП. Динамическая трансформация алгоритма ПДВП осуществляется в процессе работы ПДВП-кодера, то есть «на ходу».

Стоимостная функция $PE_{l,n}$ декомпозиции узлов $(l,n) \in E_j$ дерева ПДВП (роста структуры ПДВП) определяется как перцептуальная энтропия узла $(l,n) \in E_j$ и показывает требуемое число двоичных разрядов для кодирования звукового сигнала в частотной полосе, определяемой узлом (l,n). Функция $PE_{l,n}(l,n) \in E_j$ представляет собой функ-

цию перцептуальной энтропии Джонстона [10], однако, в отличие от известных решений, вычисляемую для действительных коэффициентов и в вейвлет-области для текущего дерева E_j ПДВП:

$$PE_{l,n} = \sum_{k=0}^{K_{l,n}-1} \log_2(2[\mathrm{n}\, int(SMR_{l,n,k})]+1), \left[\frac{\text{бит}}{(l,n)}\right],$$

где $(l,n) \in E_j$, а $k \in Z$, SMR — отношение среднеквадратического значения вейвлет-коэффициентов $X_{l,n,k}$ в полосе узла (l,n) дерева E_j к соответствующему маскирующему порогу $T_{l,n}$, равномерно распределенному между $K_{l,n}$ коэффициентами $X_{l,n,k}$, $k = \overline{1, K_{l,n}}$ узла (l,n), определяется следующим образом:

$$SMR_{l,n,k} = \left|X_{l,n,k}\right| / \sqrt{12 \cdot T_{l,n} / \mathrm{K}_{l,n}}.$$

где знаменатель $\sqrt{12 \cdot T_{l,n} / \mathrm{K}_{l,n}}$ представляет собой максимальный шаг квантователя $\Delta_{l,n}$ вейвлет-коэффициентов в узле $(l,n) \in E_j$, а величина $SMR_{l,n,k}$ задает минимальное количество уровней квантования.

Меру информативности структуры дерева ПДВП предлагается конструировать следующим образом:

$$H_{E_i} = -\sum_{\forall (l,n) \in E_i} \sum_k \frac{\left|X_{l,n,k}\right|}{\sum_{\forall (l,n) \in E_i} \left|X_{l,n,k}\right|} \ln\left(\frac{\left|X_{l,n,k}\right|}{\sum_{\forall (l,n) \in E_i} \left|X_{l,n,k}\right|}\right),$$

где $X_{l,n,k} \in (l,n)$ — коэффициенты узла (l,n) дерева E_i, $(l,n) \in E_j$, $k \in \mathbb{Z}, i = 1...8$. Данная стоимостная функция характеризует изменение во времени информативности дерева E_i ПДВП, отсюда и название временная энтропия вейвлет-коэффициентов H_E (wavelet time entropy), и представляет собой сумму энтропий $H_{l,n}$ вейвлет-коэффициентов $X_{l,n,k}$ в узлах (l,n) дерева E_i:

$$H_{l,n} = -\sum_k \frac{\left|X_{l,n,k}\right|}{\sum_{(l,n) \in E_i} \left|X_{l,n,k}\right|} \ln\left(\frac{\left|X_{l,n,k}\right|}{\sum_{(l,n) \in E_i} \left|X_{l,n,k}\right|}\right), \qquad (10)$$

где $X_{l,n,k} \in (l,n)$ — коэффициенты узла (l,n) дерева E_i, $(l,n) \in E_j$, $k \in \mathbb{Z}, i = 1...8$. Правило отбора наилучшей декомпозиции для узла базируется на выборе той декомпозиции значение энтропии вейвлет-коэффициентов, которой будет меньше. Минимизация стоимостной функции H_E ведет к сокращению неопределенности и соответственно к увеличению информативности дерева ПДВП, описывающего входной сигнал. Декомпозиция ПДВП, то есть «рост» дерева преобразования или, другими словами, динамическая трансформация частотно-временного плана, может осуществляться на основании следующего алгоритма [22, 26].

Алгоритм 1. *Алгоритм динамической трансформации частотно-временного плана на основе перцептуального критерия и оценки временной энтропии вейвлет-коэффициентов.*

Пусть решение о декомпозиции узла (l,n) дерева E_i ПДВП будет обозначаться как $split(l,n)$, где l — уровень декомпозиции, то есть масштабный уровень преобразования, а n есть n-й узел на уровне l. Пусть текущий узел (предок) будет (l,n), а его потомки определяются как $(l+1,2n)$ и $(l+1,2n+1)$, где $l = 0,1,2,3...$, $n = 0,1,2,3....$

Шаг 1. Пусть $l = 0$, $split(l,n) = YES$, то есть задан корневой узел $(0,0)$ дерева преобразования $E_0 = E_j$, где $j = 0$ — входной фрейм звукового сигнала, перцептуальная энтропия которого равна $PE_{0,0}$ и информативность дерева равна H_{E_0}. $PE_{l,2,n}$

Шаг 2. Осуществляется декомпозиция узлов предков (l,n) входного сигнала на основе банка из двух ортонормальных вейвлет-фильтров с временным сдвигом 2^0 и без него 2^1.

Шаг 3. Вычисление информативности $H_{l,n}$ в каждом узле $(l,n,2^0)$ и $(l,n,2^1)$. Выбор узла удовлетворяющего требованию:

ЕСЛИ $H_{l,n,2^0} > H_{l,n,2^1}$,

ТОГДА $(l,n) = (l,n,2^1)$

ИНАЧЕ $(l,n) = (l,n,2^0)$

Шаг 4. $l = l+1, j = j+1.$

ЕСЛИ $l-1 >$ максимального масштабного уровня предельного дерева $CB-WPD$,

ТОГДА STOP — конец роста дерева ПДВП.

Шаг 5. Вычисляется перцептуальная энтропия $PE_{l,2n}$ и $PE_{l,2n+1}$ в узлах декомпозиции $(l,2n)$ и $(l,2n+1)$ соответственно для всех n.

Шаг 6. Для каждого узла n уровня l рост дерева E_i ПДВП осуществляется следующим образом:

ЕСЛИ $PE_{l,n} \geq PE_{l+1,2n} + PE_{l+1,2n+1}$,

ТОГДА $split(l,n) = YES$,

ИНАЧЕ $split(l,n) = NO$,

Шаг 7. Оценивается информативность H_{E_j} дерева E_i:

ЕСЛИ $H_{E_j} > H_{E_{j-1}}$,

ТОГДА STOP — конец роста дерева, результирующее дерево ПДВП E_{j-1}.

Шаг 8. Переход к шагу 2.

Достоинством алгоритма 1 является то, что рост дерева осуществляется сверху вниз, без возвратов на меньшие масштабные уровни преобразования и необходимости построения полного дерева ПДВП [26], что соответствует концепции обработки сигналов в реальном времени [27]. Данный алгоритм роста дерева ПДВП позволяет определить субоптимальную структуру декомпозиции ПДВП при минимальном числе бит на отсчет звукового сигнала без воспринимаемых на слух искажений, вносимых в процессе кодирования входного сигнала. На рисунке 9 выполнена наглядная иллюстрация конструирования дерева ПДВП на уровне $l = 2$.

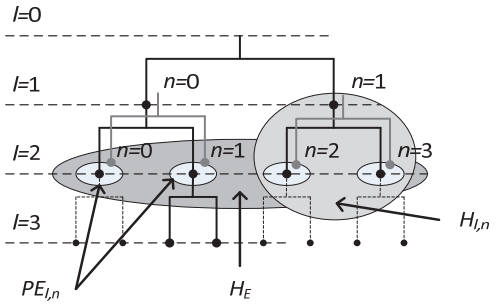


Рис. 9. Иллюстрация конструирования эффективного дерева ПДВП

Процесс роста дерева ПДВП E, согласно алгоритму динамической трансформации частотно-временного плана, и формируемый им частотно-временной план в соответствии со структурой дерева ПДВ E для некоторого фрейма входного сигнала показан на рисунке 10.

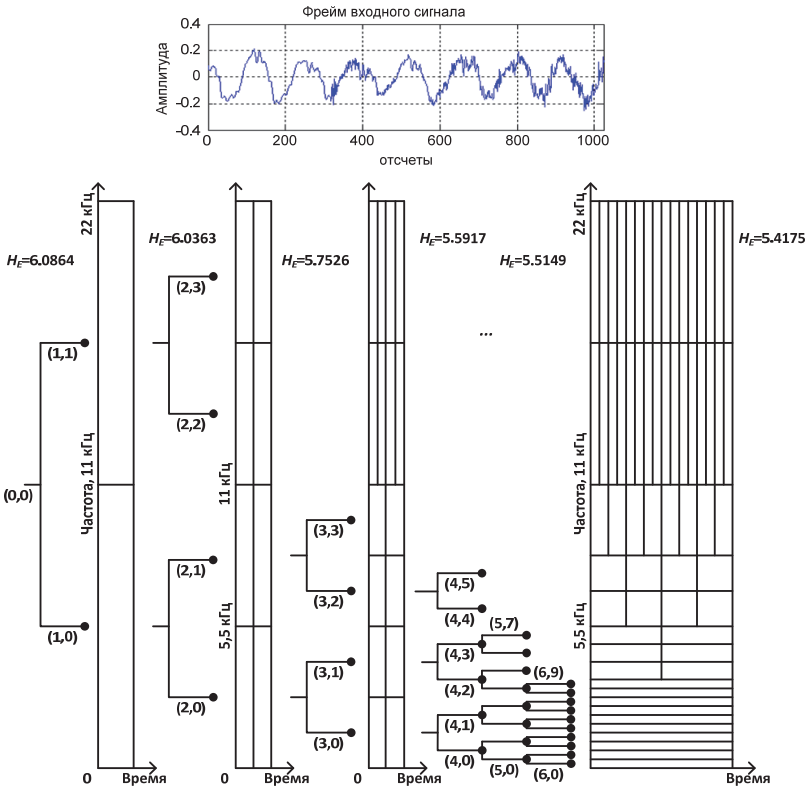


Рис. 10. Формирование дерева ПДВП и частотно-временного плана

Здесь на каждом масштабном уровне l дерева ПДВП E определяется его информативность (частотно-временного разрешения) H_E, и оценивается перцептуальная энтропия $PE_{l,n}$ для каждого узла n уровня l дерева ПДВП E. В соответствии с алгоритмом принимается решение о дальнейшем росте дерева ПДВП или о нахождении субоптимального частотно-временного разрешения для входного фрейма аудиосигнала в соответствии со структурой дерева ПДВП. Дерево $E_{1,6}$ наиболее полно

удовлетворяет требованиям обработки, то есть значение меры информативности H_{E_j} минимально. По мере увеличения декомпозиции дерева ПДВП порог маскирования — порог едва различимых искажений приближается к такому же порогу, который был построен для предельного дерева преобразования $CB-WPD$ (рисунок 11).

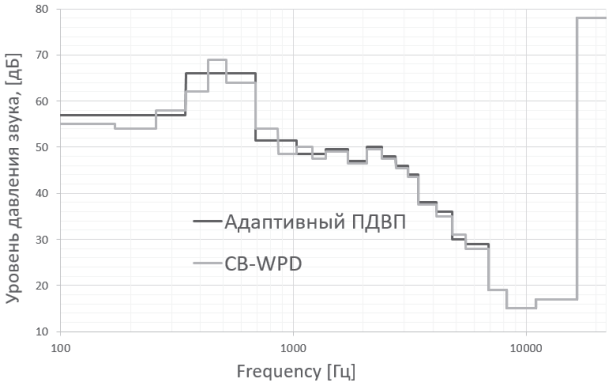


Рис. 11. Порог маскирования аудиосигнала

4. Параметрическое описание звуковых сигналов на основе согласованной подгонки.

4.1. Разреженная аппроксимация сигнала с полным ПДВП-словарем вейвлет-коэффициентов. Использование частотно-временного преобразования на основе алгоритма согласованной подгонки со словарем частотно-временных функций позволяет получить разреженную частотно-временную аппроксимацию аудиосигналов, а следовательно, лучшую, чем на основе вейвлет-преобразования и ПДВП локализацию нестационарности в сигнале [14, 31]. Это в конечном итоге позволит уменьшить цифровой поток в схеме компрессии [12, 13, 19].

Положим, что $D=(g_\gamma)_{\gamma\in\Gamma}$ — есть полный словарь частотно-временных функций, в дальнейшем называемых атомами, относительно которых параметризуется частотно-временные функции, например, масштабный фактор, частота, время индексируются $\gamma\in\Gamma$, где Γ — случайное множество индексов. Пусть дана некоторая частотно-временная функция g_γ, параметризированная относительно индекса γ, то

наилучшая возможная аппроксимация $x(n)$ получается как ортогональная проекция аудио сигнала $x(n)$ на стянутое подпространство частотно-временной функции g_γ. Декомпозиция аудиосигнала может быть представлена как $x(n) = \left\langle x(n), g_\gamma \right\rangle \cdot g_\gamma + r(\mathrm{n})$, где $r(\mathrm{n})$ — сигнал-остаток после вычитания проекции $\left\langle x(n), g_\gamma \right\rangle \cdot g_\gamma$. На основании ортогональности $r(\mathrm{n})$ и g_γ, следует, что $\|x(n)\|^2 = \left|\left\langle x(n), g_\gamma \right\rangle\right|^2 + \|r(n)\|^2$. Данная декомпозиция делается для любого и каждого элемента словаря и лучшее согласование будет найдено путем выбора элемента g_γ, для которого $\|r(\mathrm{n})\|$ минимальна, или, что эквивалентно, для которого $\left|\left\langle x(n), g_\gamma \right\rangle\right|$ — максимальный. Математически это можно сформулировать следующим образом $\gamma' = \arg\sup_{\gamma \in \Gamma} = \left|\left\langle x(n), g_\gamma \right\rangle\right|$. Таким образом, сигнал $x(n)$ проецируется на избыточный словарь частотно-временных функций со всеми возможными комбинациями масштабов, переходных и модуляционных параметров. Когда $x(n)$ реальный и дискретный, как аудиосигнал, тогда используется словарь реальных и дискретных функций. Благодаря избыточности словаря это дает возможность максимально гибко выбрать лучшую частотно-временную функцию для локальной структуры сигнала (локальная оптимизация). Максимальная гибкость модели сигнала позволяет осуществлять компактную аппроксимацию сигнала с максимально возможной точностью минимально возможным числом частотно-временных функций.

Структура алгоритма разреженной аппроксимации, используя алгоритм согласованной подгонки со словарем на основе полного дерева ПДВП, показана на рисунке 12 [31]. Для N отсчетов входного сигнала словарь D содержит N векторов $g_\gamma \in D$, где $\gamma = (l, n, k)$ индекс вектора для каждого уровня $0 < l, \log_2(N)$ узла $0 < n < 2^l$ и коэффициента $0 < k < 2^l N$ декомпозиции. Каждый вектор имеет частотно-временную локализацию схожую с дискретной функцией окна, которая увеличивается на 2^l и имеет центр в $2^l(k+1/2)$.

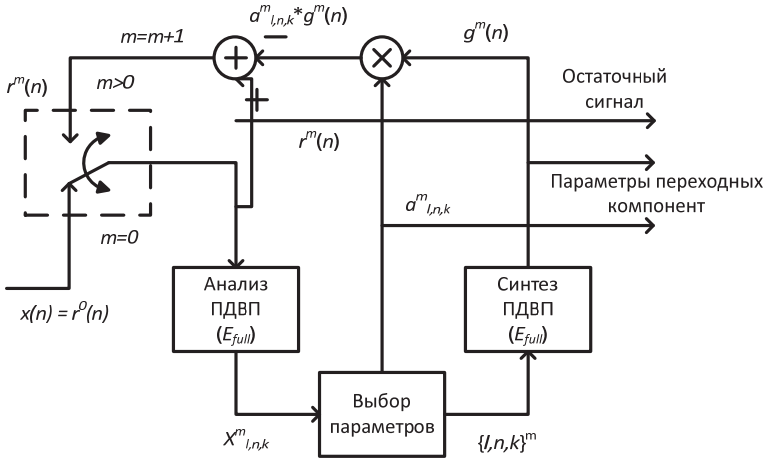


Рис. 12. Структура алгоритма согласованной подгонки со словарем, построенным на основе полного дерева ПДВП

Алгоритм согласованной подгонки реализуется итерационным повторением следующих шагов:

Шаг 1. Декомпозиция сигнала-остатка $r^m(n)$ банком фильтров на основе полного дерева ПДВП E_{full}, для $m = 1$ итерации $r^m(n) = x(n)$;

Шаг 2. Выбор наиболее значимого вейвлет-коэффициента X_γ, то есть коэффициента с абсолютным максимальным значением весового коэффициента a_γ;

Шаг 3. Наиболее значимому вейвлет-коэффициенту X_γ ставиться в соответствие атом g_γ из словаря D;

Шаг 4. Формирование результирующего вектора $g^m_\gamma(n)$ декомпозиции выполняется при помощи обратного ПДВП преобразования на основании структуры дерева ПДВП E_{full};

Шаг 5. Получение результирующего сигнала, как умножение результирующего вектора $g^m_\gamma(n)$ на весовой коэффициент a^m_γ;

Шаг 6. Получение сигнала-остатка $r^{m+1}(n)$ путем вычитания результирующего сигнала $a^m_\gamma \cdot g^m_\gamma(n)$ из остаточного сигнала $r^m(n)$.

Процесс повторяется и на следующей итерации алгоритма входным сигналом будет сигнал-остаток с предыдущей итерации. На m-й итерации сигнал-остаток $r^m(n)$ вычисляется как:

$$r^m(n) = \begin{cases} x(n) & m = 0 \\ r^{m+1}(n) + a_\gamma^m \cdot g_\gamma^m(n) & m \neq 0 \end{cases}, \quad (1)$$

где a_γ^m весовой коэффициент оптимального вектора $g_\gamma^m(n)$ на m-й итерации, а γ индекс словаря D на m-й итерации. Оптимальным вектором считается тот вектор, у которого получается наибольшее значение произведения с сигналом-остатком $\langle x(n), g_\gamma \rangle$.

Минимизация вычислительной сложности метода напрямую связана также со структурой и размером словаря. Количество атомов словаря D находится в прямой зависимости от числа уровней дерева ПДВП и длинны анализируемого фрейма входного сигнала. С точки зрения, например, перцептуального восприятия акустической информации человеком, наиболее критичные компоненты сигнала расположены в области нижних, нежели в области верхних частот, что сказывается на качестве выбора перцептуально значимых полос для анализа.

4.2. Согласованная подгонка с перцептуально оптимизированным словарем частотно-временных функций. В [32] предлагается задачу параметрического анализа аудиосигналов решать на основе разреженной аппроксимации с перцептуально оптимизированным ПДВП-словарем вейвлет-коэффициентов. Алгоритм согласованной подгонки требует значительных вычислительных затрат от итерации к итерации при формировании субоптимального решения, которое в конце концов может быть не применимо в некоторых приложениях обработки звука и речи. Задача поиска оптимального словаря D, построенного на основе ПДВП, сводится к поиску лучшей структуры дерева декомпозиции ПДВП [30]. Формирование оптимального набора частотно-временных функций в словаре выполняется на основе применения перцептуально адаптированного к текущему фрейму входного сигнала ПДВП, что позволяет уменьшить размер формируемого словаря и сделать его динамическим, то есть зависимым от структуры аудиосигнала [25, 32].

Структурная схема параметрического анализа аудиосигналов на основе алгоритма согласованной подгонки с перцептуально адаптированным к фрейму входного сигнала ПДВП словарем состоит из двух частей:

— первая часть — схема оптимизации структуры дерева ПДВП под фрейм входного сигнала согласно алгоритму динамической трансформации частотно-временного плана входного сигнала на основе перцертуальных критериев [26];

— вторая часть представляет собой схему модифицированного алгоритма согласованной подгонки с перцептуально-оптимизированным ПДВП словарем.

Структура модифицированного алгоритма согласованной подгонки с динамически оптимизированным словарем показана на рисунке 13.

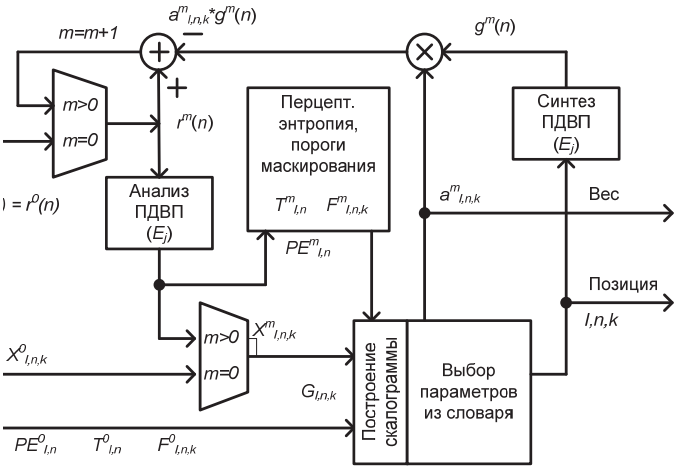


Рис. 13. Схема параметрического анализа аудиосигналов на основе разреженной аппроксимации с перцептуально оптимизируемым ПДВП-словарем

На итерации $m = 0$ коммутация выполнена так, чтобы остаточный сигнал $r^m(n)$ стал равным фрейму входного сигнала $x(n)$, а вейвлет-коэффициенты $X^0_{l,n,k}$ попали напрямую в блок скалярного произведения и скалограммы возбуждения соответственно. Первоначальное значение перцептуальной энтропии $PE^0_{l,n}$, порогов маскирования $T^0_{l,n}$ и временных маскеров $F^0_{l,n,k}$ берется из схемы декомпозиции входного сигнала адаптивным ПДВП. В блоке скалярного произведения и скалограммы возбуждения выполняется формирование скалограммы возбуждения для текущего входного фрейма сигнала $G^0_{l,n,k}$, отбор по ней перцептуально значимого вейвлет-коэффициента, формирование весового параметра атома

$\alpha^0_{l,n,k}$ и позиции вейвлет-коэффициента. В блоке обратного ПДВП выполняется синтез единичной атомарной функции $g^0(n)$ на основании знания позиции вейвлет-коэффициента в структуре дерева E_j ПДВП, затем умножение ее на весовой коэффициент $\alpha^0_{l,n,k}$. Остаточный сигнала, он же входной сигнал для следующей итерации, формируется вычитанием результирующего сигнала $\left\langle \alpha^0_{l,n,k} \cdot g^0(n) \right\rangle$ из сигнала $r^0(n)$. На каждой последующей итерации $m > 0$ пороги маскирования $T^m_{l,n}$ и временные маскеры $T^m_{l,n,k}$ вычисляются для формирования скалограммы возбуждения $G^m_{l,n,k}$, на основании которой ведется отбор перцептуально-значимых компонент сигнала [17]. Работа схемы параметрического анализа аудиосигналов на основе разреженной аппроксимации с перцептуально оптимизируемым ПДВП-словарем выполняется согласно следующему алгоритму [32].

Алгоритм 2. *Алгоритм согласованной подгонки с динамически оптимизируемым ПДВП словарем на основе психоакустического критерия.*

Исходные данные алгоритма согласованной подгонки: E_j — оптимизированная структура ПДВП для фрейма входного сигнала $x(n)$; $T_{l,n}$ — порог маскирования для каждого оконечного узла $(l,n) \in E_j$ ПДВП; $F_{l,n}$ — временный маскер для каждого оконечного узла $(l,n) \in E_j$ ПДВП; $G_{l,n,k}$ — скалограмма возбуждения соответствующая фрейму входного сигнала $x(n)$.

Шаг 1. Установить номер итерации $m = 0$.

Шаг 2. Разместить $G_{l,n,k}$ и установить $G_{l,n,k} = 0$ для всех l,n,k в соответствии со структурой дерева ПДВП E_j.

Шаг 3. Вычислить $PE^m_{l,n}$ для всех узлов (l,n), используя $T_{l,n}$;

ЕСЛИ $PE^m_{l,n} == 0 \forall (l,n,k) \in E_j$

ТОГДА СТОП;

ЕСЛИ $PE^m_{l,n} == 0$,

ТОГДА $X_{l,n}^{m} = 0$ для $k = \{0, K_{l,n} - 1\}$ узла (l,n).

Шаг 4. Выбрать из $X_{l,n,k}^{m}$ значимые коэффициенты $X_{l,n,k}^{*m}$, имеющие наибольший вес возбуждения.

Шаг 5. Создать скалограмму возбуждения, соответствующую моделируемому сигналу, используя $T_{l,n}$ и $F_{l,n}^{m-1}$ для выполненной итерации и каждого нового значимого коэффициента $X_{l,n,k}^{*m}$.

Шаг 6. Выбрать вес $\alpha_{l,n,k}^{m} = X_{l,n,k}^{*m}$, который улучшает соответствие между скалограммами (исходной и моделируемой).

Шаг 7. Получить позицию коэффициента в структуре дерева ПДВП: $l^* = l$, $n^* = n$, $k^* = k$.

Шаг 8. Установить 1 в позиции $(l^*, n^*, k^*) : G_{l^*,n^*,k^*} = 1.$

Шаг 9. Синтезировать атом $g^m(n)$ из G_{l^*,n^*,k^*}, используя обратный ПДВП со структорой дерева E_j, соответствующей ПДВП-словарю.

Шаг 10. Вычислить сигнал-остаток $r^m(n)$ из $g^m(n)$ и $\alpha_{l,n,k}^{m}$ в соответствии с (11).

Шаг 11. Применить оптимизированный на основе текущего фрейма ПДВП со структурой дерева E_j к сигналу-остатку $r^m(n)$.

Шаг 12. Увеличить номер итерации $m = m + 1$.

Шаг 13. Перейти к шагу 2.

Число итераций алгоритма 2 определяется числом перцептуально значимых коэффициентов ПДВП в сигнале остатке. В алгоритм введена процедура расчета перцептуальной энтропии $PE_{l,n}$ для каждой полос (l,n), на основании которой ведется селекция перцептуально-значимых частотных полос. Знание не только частотной составляющей сигнала, то есть порогов маскирования $T_{l,n}$, но и временного маскера $F_{l,n,k}$ позволяет обеспечить оптимальный выбор значимых компонент сигнала. Данный подход позволит уменьшить число итераций алгоритма согласованной подгонки и производить выбор только значимых компонент сигнала, что будет соответствовать оптимальным векторам

декомпозиции. Однако поиск наилучшей аппроксимации анализируемого входного сигнала $x(n)$ векторами $g_{l,n,k}$ из словаря D, построенного на основе перцептуально-адаптированного ПДВП, сложная в вычислительном плане задача.

На рисунке 14 показан пример аппроксимации аудиосигнала на основе предложенного метода. Функция Db20 из семейства Добеши с 40 коэффициентами использовалась для анализа входного сигнала в схеме перцептуально адаптированного ПДВП. Скалограммы возбуждения для входного сигнала (рисунок 14а) и синтезированного сигнала на основе 5 атомов (рисунок 14б) приведены соответственно на рисунке 15а и рисунке 15б, где перцептуально значимые компоненты показаны как положительные элементы.

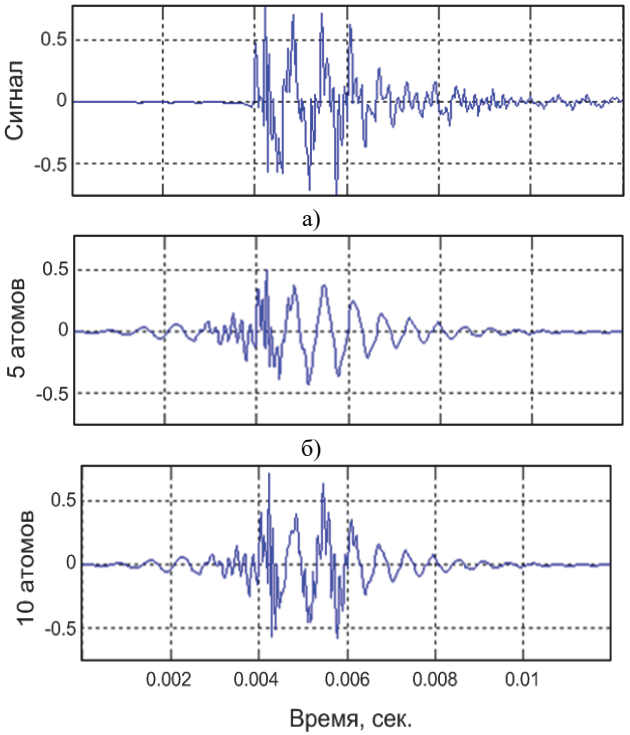
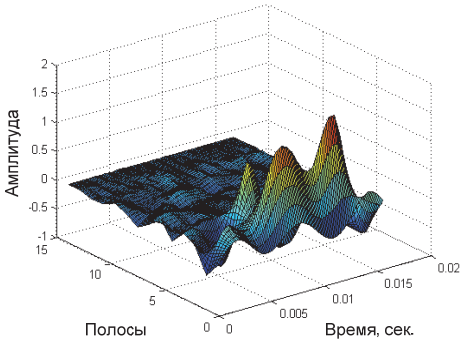
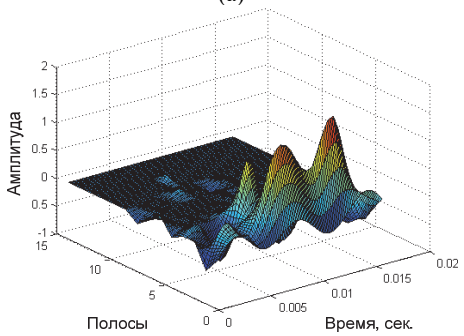


Рис. 14. Пример разреженной аппроксимации с перцептуально оптимизированным ПДВП словарем вейвлет-функций: (а) исходный сигнал; (б) 5 атомов; (в) 10 атомов



(а)



(б)

Рис. 15. Скалограммы возбуждения: (а) входного сигнала;
(б) синтезированного сигнала, на основе 5 атомов

Сравнение сходимости алгоритма согласованной подгонки на основе трех различных словарей аппроксимирующих функций показано на рисунке 16.

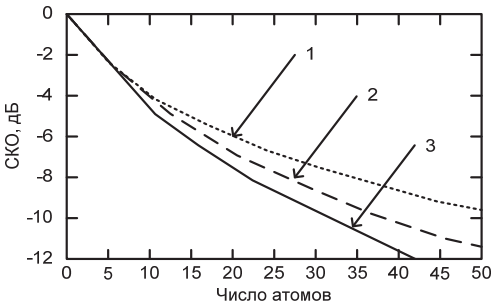


Рис. 16. Среднеквадратичное отклонение ошибки аппроксимации:
1 – затухающие синусоиды; 2 – вейвлет-пакет с избыточным словарем;

Алгоритм 2 согласованной подгонки на базе перцептуально оптимизированного ПДВП-словаря вейвлет-функций [32] имеет меньшее среднеквадратичное отклонение ошибки аппроксимации, по сравнению с алгоритмами на основе ПДВП-словаря с полным набором вейвлет-функций [31] и с использованием затухающих синусоид [33].

Таким образом, в алгоритме 2 формирование оптимального набора частотно-временных функций в словаре выполняется на основе применения перцептуально адаптированного к текущему фрейму входного сигнала ПДВП, что позволяет уменьшить размер формируемого словаря и сделать его динамическим, то есть зависимым от структуры фрейма аудиосигнала. Преимуществом алгоритма 2 разреженной аппроксимации с перцептуально оптимизированным ПДВП-словарем вейвлет-коэффициентов является высокая скорость сходимости и минимизация перцептуальных искажений, обусловленная построением перцептуально оптимизированного частотно-временного плана соответствующей декомпозиции вейвлет-пакета текущего сигнала-остатка для выбора оптимальных частотно-временных функции для каждой итерации подгонки. Более того, существует точный психоакустический критерий остановки описанной процедуры.

5. Параметрический аудиоречевой кодер с разреженной аппроксимацией сигнала и перцептуально-оптимизированным словарем частотно-временных функций.

5.1. Структура параметрического аудиоречевого кодера. Структура параметрического перцептуального аудиоречевого кодера показана на рисунке 17, который состоит из блока динамической трансформации частотно-временного плана на базе алгоритма 1 (отмечен пунктирной линией) и блока разреженной аппроксимации сигнала на основе алгоритма 2, согласованной подгонки с динамически оптимизированным ПДВП словарем частотно-временных функций (серая область, выделенная пунктирной линией). Результатом адаптивного анализа являются вейвлет-коэффициенты $X_{l,n,k}$, трехмерная скалограмма возбуждения обрабатываемого фрейма сигнала, построенная на их основе, и структура дерева ПДВП E_j. Перцептуально значимые вейвлет-коэффициенты $X_{l,n,k}$, определенные в алгоритме согласованной подгонки и их позиции в дереве ПДВП из блока разреженной аппроксимации сигнала поступают в блок квантования и кодирования. В [22] разработано правило оптимального распределения битов при квантовании

вейвлет-коэффициентов в кодере с учетом реконструкции сигнала в декодере при переменном коэффициенте децимации в каналах банков фильтров кодера и декодера. Так, «эффект просачивания» энергии шума квантования в смежные полосы банка фильтров декодера не может пренебрегаться без внесения ущерба в качество восстановленного аудиосигнала. Показано в [22], что выбирая достаточно большой порядок (20-й и более) вейвлет-функции (высокую частотную избирательность канальных фильтров), можно ограничиться одинаковыми фильтрами как в анализирующем (кодере), так и синтезирующем (декодере) банках фильтров.

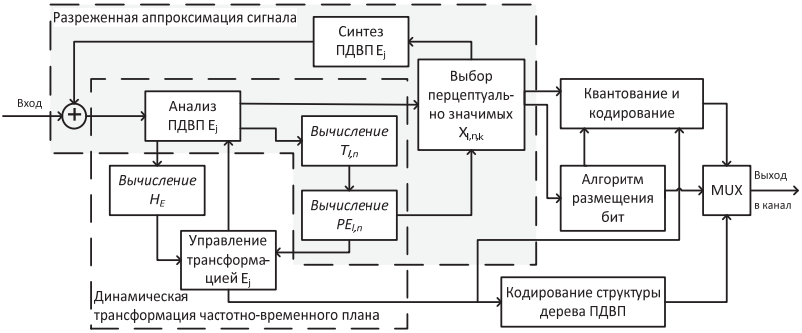


Рис. 17. Структура параметрического аудиоречевого кодера

Структура закодированных перцептуально значимых компонент — вейвлет-коэффициентов $X_{l,n,k}$ такова, что на стороне декодера после деквантования, параметры записываются в уже подготовленную структуру критического дерева ПДВП – *CB WPD* и выполняется синтез сигнала на основе обратного ПДВП – *CB WPD*. Структура параметрического аудиоречевого декодера показана на рисунке 18. Работа декодера выполняется в следующем порядке: входная информация декодируется и восстанавливается в блоке декодирования и восстановления. Полученные параметры содержат значения перцептуально значимых компонент — вейвлет-коэффициентов $X_{l,n,k}$ и информацию о их местоположении в структуре критического дерева ПДВП – *CB WPD*. Синтезированные фреймы сигнала умножаются на треугольное окно и складываются для формирования реконструированного сигнала $\hat{x}(n)$.

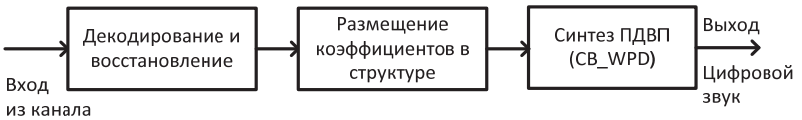


Рис. 18. Структура параметрического аудиоречевого декодера

Масштабирование цифрового потока осуществляется выбором условия остановки разреженной аппроксимации сигнала на основе алгоритма 2 согласованной подгонки с динамически оптимизированным ПДВП словарем частотно-временных функций. В данном кодере используются два варианта: фиксированное число атомов, либо определенный порог отобранной алгоритмом согласованной подгонки энергии или по уровню энергии остаточного сигнала $r^m(n)$. Для условия остановки алгоритма в виде фиксированного числа атомов задается определенное число итераций m, по достижению которого прекращается работа алгоритма кодирования, и реконструкция сигнала осуществляется на основе отобранных атомов. Это дает дополнительные возможности масштабирования цифрового потока кодера, так как количество атомов для каждого фрейма может быть различным, следовательно, есть возможность динамического добавления дополнительных пакетов атомов от фрейма к фрейму при наличии возможности канала передачи данных либо по запросу потребителя. Таким образом, этот факт дает возможность выстроить многоуровневую структуру аудиокодера относительно качества реконструированного сигнала. В проводимых экспериментах, за базовый уровень, при котором искажения восстановленного сигнала не являются раздражающими (на основе объективной оценки качества), бралось значение в 200 атомов. Добавляя к базовому количеству атомов дополнительные пакеты (в тестах — 50 атомов) достигается увеличение качества при росте скорости цифрового потока.

В силу того, что структура кодера (рисунок 17) работает по принципу анализа через синтез, и на каждой итерации производятся вычисления сигнала остатка $r^m(n)$ текущего фрейма, то без дополнительных изменений схемы кодирования можно отслеживать затухание энергии остаточного сигнала. При достижении определенного уровня процедуру согласованной подгонки можно остановить. Это дает возможность динамически изменять скорость цифрового потока в зависимости от структуры входных данных и отобранной для передачи информации.

5.2. Экспериментальные результаты. Для объективной оценки качества восстановленного аудиоречевого сигнала была использована модель $PEMO-Q$ [34]. Данная методика имеет низкую степень специализации относительно оцениваемого сигнала, что хорошо подходит для оценки качества различных по своей природе аудио образцов. Шкала оценки (*Objective Difference Grade* — ODG) в зависимости от степени искажения выходного сигнала формируется следующим образом: не воспринимаемое искажение (*«imperceptible»*) = 0; воспринимаемое, но не раздражающее (*«perceptible but not annoying»*) = –1.0; не-

много раздражающее (*«slightly annoying»*) = –2.0; раздражающее (*«annoying»*) = –3.0; очень раздражающее (*«very annoying»*) = –1.0. Входной тестовой последовательностью служили одноканальные образцы звуковых сигналов (таблица 1). Частота дискретизации 44.1 кГц, разрядность представления отсчетов сигнала – 16 бит.

Таблица 1. Тестовые образцы

Образец	Описание	Образец	Описание
$\boldsymbol{es01}$	Вокал (Suzan Vega)	$si01$	Клавесин
$\boldsymbol{es02}$	Речь на немецком языке	$si02$	Кастаньеты
$\boldsymbol{es03}$	Речь на английском языке	$si03$	Pitch pipe
$\boldsymbol{sc01}$	Соло на трубе и оркестр	$sm01$	Волынка
$\boldsymbol{sc02}$	Оркестровая композиция	$sm02$	Металлофон
$\boldsymbol{sc03}$	Современная поп-музыка	$sm03$	Plucked strings

В ходе экспериментов было рассчитано, что бюджет бит для варианта с 200 атомами и учетом затрат на кодирование местоположения атомов в структуре критического дерева ПДВП $— CB — WPD$ составляет 37 кбит/с, а каждые дополнительные 50 атомов увеличивают битрейт на 8,6 кбит/с. Объективные оценки качества ODG каждого испытуемого образца показаны на рисунке 19 [18].

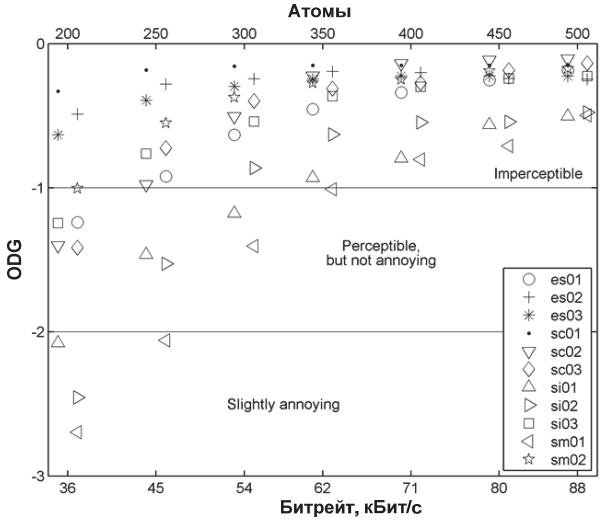


Рис. 19. Распределение объективных оценок качества ODG

Как видно из рисунка 19, ни для одного из исследуемых образцов объективная оценка ODG не находится ниже *«slightly annoying»*. На самом низком битрейте (200 атомов) только три образца в зоне *«slightly annoying»*: речевые образцы ($es02$, $es03$). Речевой образец $es01$ (вокал,

Suzan Vega) имеет отметку в диапазоне *«perceptible, but not annoying»*. С ростом скорости передачи распределение объективных оценок ODG сдвигается в область *«imperceptible»*. Для 350 атомов (или 62 кбит) все образцы имеют значения ODG от -1 до 0 (только $sm01$ находится на границе -1). Применение схемы масштабируемости позволяет подтянуть оценки до *«imperceptible»* для всех тестовых образцов, изменяя число атомов на ходу для каждого кадра.

Анализ оригинального сигнала состоящего из смеси тестовых образцов $es03$, $sc01$, $si02$, $sm01$ и восстановленных сигналов для разного числа атомов разреженной аппроксимации в кодере, а также их спектрограмм (рисунок 20) показывает коррелированность результатов с объективными оценками ODG (рисунок 19). Таким образом, анализ экспериментальных результатов показывает, что предложенный универсальный аудиоречевой кодер обеспечивает хорошее качество восстановленного сигнала и эффективно работает с различными типами входного звукового содержимого.

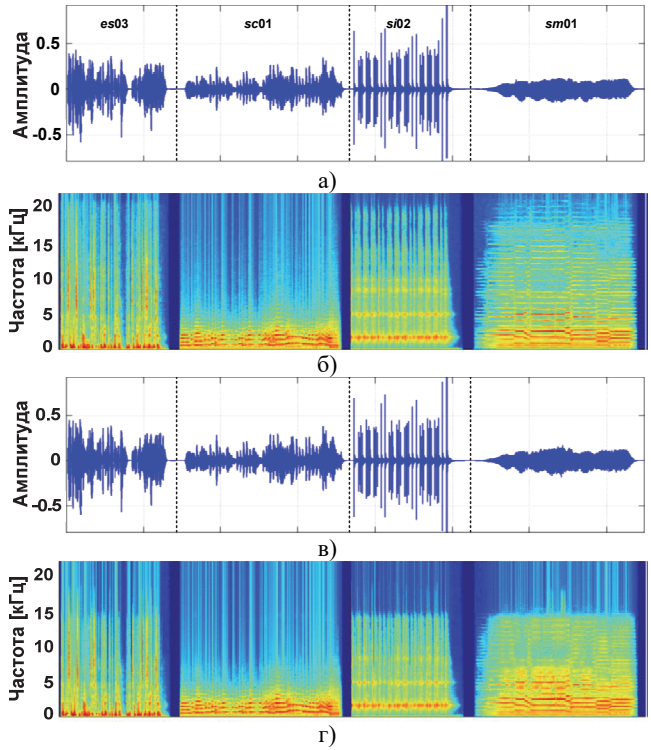


Рис. 20. Оригинальный и восстановленные сигналы для разного числа атомов разреженной аппроксимации: а, б) – оригинальный сигнал; в, г) восстановленный сигнал, число атомов 200

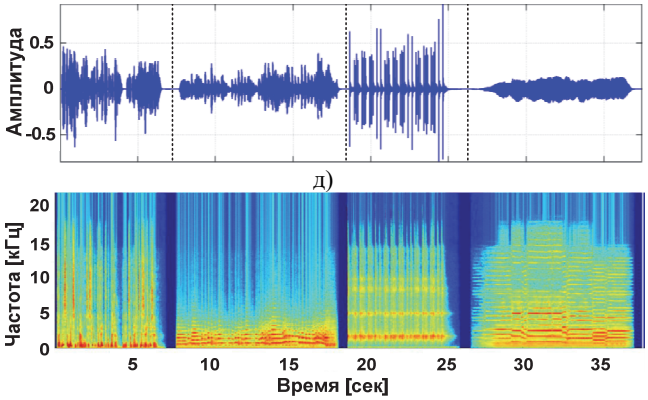


Рис. 20. Оригинальный и восстановленные сигналы для разного числа атомов разреженной аппроксимации: д, е) восстановленный сигнал, число атомов 500

Тестовые образцы (таблица 1) также были сжаты алгоритмами *Vorbis* и *Opus*[35, 36]. Результаты сравнения оценок объективного качества ODG данного масштабируемого аудиоречевого кодера (MP coding scheme) с современными кодерами (*Opus* и *Vorbis*) показано на рисунке 21 [18, 19].

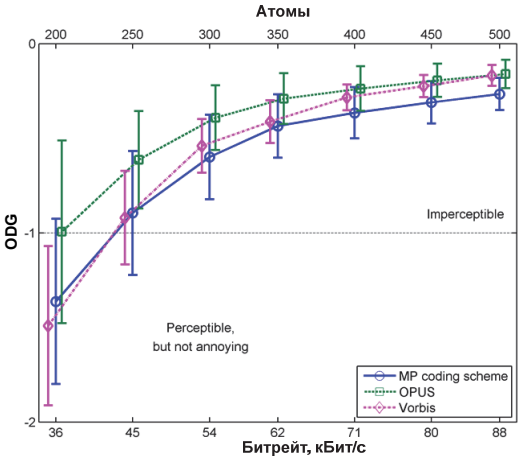


Рис. 21. Объективные оценки качества ODG кодеров MP coding scheme, *Opus* и *Vorbis*

Анализ результатов эксперимента показывает, что от 45 кбит и выше все три кодера находятся в *«imperceptible»* диапазоне. Оценка качества ODG для *Opus* (36 кбит) находится на границе *«imperceptible»* области. Оценка качества ODG для алгоритма *Vorbis* немного меньше и

попадает в область *«perceptible, but not annoying»*. Стоит отметить, что *Opus* является составным кодером, который состоит из двух различных алгоритмов для кодирования речи и музыки, в то время как разрабатываемый аудиокодер использует одну модель для работы со всеми типами входных сигналов.

6. Заключение. Два алгоритма адаптивного частотно-временного анализа звуковых сигналов предложены в работе: алгоритм динамической трансформации частотно-временного плана, позволяющий определить субоптимальную структуру декомпозиции ПДВП, достоинством которого является то, что рост дерева осуществляется сверху вниз, без возвратов на меньшие масштабные уровни преобразования и необходимости построения полного дерева ПДВП, что соответствует концепции обработки в реальном масштабе времени; алгоритм параметрического анализа сигналов на основе применения разреженной аппроксимации. Формирование оптимального набора частотно-временных функций в словаре выполняется на основе применения перцептуально адаптированного к текущему фрейму входного сигнала ПДВП, что позволяет уменьшить размер формируемого словаря и сделать его динамическим, то есть зависимым от структуры аудиосигнала. Предложено структурное решение масштабируемых перцептуальных аудиоречевых кодеров на базе ПДВП с динамической реконфигурацией частотно-временного плана: параметрический аудиоречевой кодер на основе разреженной аппроксимации с перцептуально оптимизированным словарем частотно-временных функций. При сравнении предложенного масштабируемого кодера с известными универсальными кодерами *Vorbis* и *Opus* можно сделать вывод, что часть тестовых звуковых образцов эквивалентна по качеству при равных скоростях цифрового потока, масштабирование которого, в разработанном кодере, позволяет при небольшом увеличении скорости цифрового потока добиться эквивалентного качества и для остальных образцов.

Литература

1. *Kahrs M., Brandenurg K.* Application of digital signal processing to audio and acoustics // USA Boston: Kluwer Academic Publishers. 1998. 545 p.
2. *Valin J.-M., Terriberry T. B., Montgomery C., Maxwell G.,* A high-quality speech and audio codec with less than 10-ms delay, // IEEE Transaction on audio, speech, and language processing. 2010. vol. 18. pp. 58–67.
3. *Umapathy K., Ghoraani B., Krishnan S.,* Audio signal processing using time-frequency approaches: coding, classifcation, fingerprinting, and watermarking // EURASIP Journal on Advances in Signal Processing. 2010. vol. 2010. no. 1. pp. 451695.
4. *Painter T., Spanias A.* Perceptual Coding of Digital Audio // Proceedings of IEEE. 2000. vol. 88. no. 4. pp. 451–513.
5. *Brandenburg K.* Introduction to perceptual coding // Collected Papers on Digital Audio Bit-Rate Reduction. Eds. 1996. pp. 23–30.

6. *Spanias A., Painter T., Atti V.* Audio signal processing and coding // John Wiley & Sons, Inc. New Jersey. USA. 2007. 464 p.
7. *Вашкевич М.И., Азаров И.С., Петровский А.А.* Косинусно-модулированные банки фильтров с фазавым преобразованием: реализация и применение в слуховых аппаратах // Горячая линия-Телеком. Москва. 2014. 210 с.
8. *Bosi M., Goldberg R.E.* Introduction to digital audio coding and standards // Springer Science+Business Media. USA. 2003. 434 p.
9. *Wickerhauser M.V.* Adaptive Wavelet Analysis from Theory to Software // Massachusetts: A.K. Peters Ltd. 1994. 486 p.
10. *Johnston J.D.* Transform coding of audio signals using perceptual noise criteria // IEEE Transaction on Selected Areas of Communication. 1988. vol. 6. pp. 314–323.
11. *Ковалгин Ю.А., Вологдин Э.И.* Аудиотехника // Горячая линия-Телеком. 2013. 742 с.
12. *Reyes N.R., Candeas P.V.* Adaptive signal modeling based on sparse approximations for scalable parametric audio coding // IEEE Transactions on audio, speech, and language processing. 2010. vol. 18. pp. 447–460.
13. *Petrovsky Al., Herasimovich V., Petrovsky A.* Scalable parametric audio coder using sparse approximation with frame-to-frame perceptually optimized wavelet packet based dictionary // 138th AES Convention. 2015. paper 9264. 10 p.
14. *Mallat S., Zhang Z.* Matching pursuits with time-frequency dictionaries // IEEE Transaction on Signal Processing. 1993 vol. 41. no. 12. pp. 3397–3415.
15. *Chardon G., Necciari T., Balazs P.* Perceptual matching pursuit with Gabor dictionaries and time-frequency masking // Proceedings of IEEE ICASSP'2014. 2014. pp. 3126–3130.
16. *Ravelli E., Gaeul R., Daudet L.*, Matching pursuit in adaptive dictionaries for scalable audio coding // Proceedings of EUSIPCO'2008. 2008. pp. 1–5.
17. *Petrovsky Al., Azarov E., Petrovsky A.* Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding // Signal Processing. Special issue "Fourier Related Transforms for Non-Stationary Signals". 2011. vol. 91. Issue 6. pp. 1489–1504.
18. *Petrovsky Al., Herasimovich M., Petrovsky A.* Bio-inspired sparse representation of speech and audio using psychoacoustic adaptive matching pursuit // Proceedings of 18th International Conference of SPECOM 2016. 2016. pp. 156–164.
19. *Petrovsky Al., Herasimovich V., Petrovsky A.* Audio/speech coding using frame-based psychoacoustic optimized time-frequency dictionaries and its performance evaluation // IEEE conference proceedings "Signal processing: algorithms, architectures, arrangements, and applications" (SPA-2016). 2016. pp. 225–229.
20. *Burrus C.S., Gopinath R.A., Guo H.* Introduction to wavelets and wavelet transforms // N.J.: Prentice Hall. 1998. 298 p.
21. *Cohen I., Raz S., Malah D.* Orthonormal shift-invariant adaptive packet decomposition and representation // Signal Processing. 1997. vol. 57. Issue 3. pp. 251–270.
22. Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика / под ред. проф. А.А. Петровского // Минск: Бестпринт. 2009. 455 с.
23. Zwicker E., Fastl H. Psychoacoustics: Facts and Models // Berlin, Germany: Springer-Verlag. 1990. 380 p.
24. ITU-R Recommendation BS.1387, Method for Objective Measurements of Perceived Audio Quality. 1998.
25. *Petrovsky Al.* A multiresolution auditory model using adaptive WP excitation scalograms // Polska akademia nauk "Elektronika". 2008. vol. 49. no 4. pp. 65–70.
26. *Petrovsky Al., Krahe D., Petrovsky A.* Real-time wavelet packet-based low bit rate audio coding on a dynamic reconfigurable system // Proc. 138th Convention. 2003. paper 5778. 22p.

27. *Petrovsky Al., Rodionov M., Petrovsky A.* Dynamic reconfigurable on the lifting steps wavelet packet processor with frame-based psychoacoustic optimized time-frequency tiling for real-time audio applications // Design and architectures for digital signal processing. InTech. 2013. pp. 3–30.
28. *Karmakar A., Kumar A., Patney R.K.* Synthesis of an optimal wavelet based on auditory perception criterion // EURASIP Journal on Advance in Signal Processing. 2011. vol. 2011. no. 1. pp. 170927.
29. *Петровский Ал.А.* Построение психоакустической модели в области вейвлет коэффициентов для перцептуальной обработки звуковых и речевых сигналов // Научно-практический журнал «Речевые технологии». Москва. 2008. № 4. С. 61–71.
30. *Coifman R., Wickerhauser M.V.* Entropy-Based Algorithms for Best Basis Selection // IEEE Transaction on Information Theory. 1992. vol. 38. no. 2. pp. 713–718.
31. *Vera-Candeas P., Ruiz-Reyes N., Roza-Zurera M.* Transient modelling by Matching-Pursuits with a wavelet dictionary for parametric audio coding // IEEE Signal Processing Letters. 2004. vol. 11. no. 3. pp. 349–352.
32. *Petrovsky Al., Petrovsky A.* Matching pursuit algorithm with frame-based auditory optimized WP-dictionary for audio transient modeling // Polska academia nauk "Elektronika". 2008. vol. 49. no. 4. pp. 74–79.
33. *Heusdens R., Vafin R., Kleijn W.B.* Sinusoidal modeling using psychoacoustic-adaptive matching pursuits // IEEE Signal Processing Letters. 2002. vol. 9. no. 8. pp. 262–265.
34. *Huber R., Kollmeier B*PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception // IEEE Transactions on audio, speech, and language processing. 2006. vol. 14. pp. 1902–1911.
35. *Vos K., Sørensen K. V., Jensen S. S., Valin J.-M,* Voice coding with Opus // Proc. AES 135th Convention. 2013. paper 8941.10 p.
36. *Valin J.-M., Maxwell G., Terriberry T.B., Vos K*High-quality, low-delay music coding in the Opus codec // Proc. AES 135th Convention. 2013. paper 8942. 10 p.

Петровский Алексей Александрович — д-р техн. наук, доцент, руководитель группы по обработке аудио сигналов и речи в отделе мультимедийных приложений департамента исследований и разработок, ООО «Техкомпания Хуавей». Область научных интересов: цифровая обработка сигналов, системы кодирования аудиоданных, редактирования шума и устранения акустического эха, а также обработка сигналов в реальном времени для мультимедиа приложений. Число научных публикаций — 80. alexey@petrovsky.eu; Алтуфьевское шоссе 1/7, Москва, 127106, Россия; р.т.: +7(495)660-44-59.

Петровский Александр Александрович — д-р техн. наук, профессор, заведующий кафедрой электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники (БГУИР). Область научных интересов: цифровая обработка сигналов, ситемы кодирования аудио и видео данных, мультипроцессорные системы реального времени для мультимедиа приложений. Число научных публикаций — 662. palex@bsuir.by; ул. П. Бровки, 6, Минск, 220013, Беларусь; р.т.: +3751729323-40.

DOI 10.15622/sp.50.3

AL.A. PETROVSKY, A.A. PETROVSKY

SCALABLE SPEECH AND AUDIO CODERS BASED ON ADAPTIVE TIME-FREQUENCY SIGNAL ANALYSIS

Petrovsky, Al.A, Petrovsky A.A. **Scalable Speech and Audio Coders based on Adaptive Time-Frequency Signal Analysis.**

Abstract. The paper discusses the methods of perceptual sub-band audio signal processing with the dynamic time-frequency map transformation based on the discrete wavelet packet (DWPT) transform. The advantage of these methods is a top-down construction of the DWPT tree without returning to smaller scale levels of decomposition and needing to build a complete DWPT tree. This corresponds to the concept of scalable audio/speech coders implementation in real time. The objective quality assessment of the proposed coders, based on techniques PEMO-Q, and comparisons with the widespread encoders *Opus* and *Vorbis* are given. It shows that the reconstructed signal complies with ITU-R PEAQ at a high compression ratio of up to 18 times or more, does not contain artifacts: noise-to-mask ratio NMR_{total} is less ≈ -9 dB.

Keywords: scalable audio/speech coder, wavelet packet, matching pursuit

Petrovsky Alexey Aleksandrovich — Ph.D., Dr. Sci., associate professor, head of the speech and audio processing group of the media engineering department, Russian Research Center of Huawei Technologies. Research interests: acoustic signal processing, such as wideband speech and audio processing, perceptual coding, psychoacoustics, noise reduction and echo cancellation, and real-time signal processing. The number of publications — 80. alexey@petrovsky.eu; 1/7, Altufevskoe shosse, Moscow, 127106, Russia; office phone: +7(495)660-44-59.

Petrovsky Alexander Alexandrovich — Ph.D., Dr. Sci., professor, head of the Computer Engineering Department, Belarusian State University of Informatics and Radioelectronics (BSUIR). Research interests: digital signal processing, speech and audio coding, multiprocessor real-time systems for multimedia applications. The number of publications — 662. palex@bsuir.by; 6, P.Brovky str. Minsk, 220013, Republic of Belarus; office phone: +37517293-23-40.

References

1. Kahrs M., Brandenurg K. Application of digital signal processing to audio and acoustics. USA Boston: Kluwer Academic Publishers. 1998. 545 p.
2. Valin J.-M., Terriberry T. B., Montgomery C., Maxwell G., A high-quality speech and audio codec with less than 10-ms delay. *IEEE Transaction on audio, speech, and language processing*. 2010. vol. 18. pp. 58–67.
3. Umapathy K., Ghoraani B., Krishnan S., Audio signal processing using time-frequency approaches: coding, classifcation, fingerprinting, and watermarking. *EURASIP Journal on Advances in Signal Processing*. 2010. vol. 2010. no. 1. pp. 451695.
4. Painter T., Spanias A. Perceptual Coding of Digital Audio. Proceedings of IEEE. 2000. vol. 88. no. 4. pp. 451–513.
5. Brandenburg K. Introduction to perceptual coding. Collected Papers on Digital Audio Bit-Rate Reduction. Eds. 1996. pp. 23–30.
6. Spanias A., Painter T., Atti V. Audio signal processing and coding. John Wiley & Sons, Inc. New Jersey. USA. 2007. 464 p.
7. Vashkevich M.I., Azarov I.S., Petrovsky A.A. *Kosinusno-modulirovanie banki filtrov s phasovim preobrasovaniem: realizatsia i primenenie v sluchovych apparatach* [Cosine

modulated filter banks with phase transform: implementation and application to the hearing aids]. Moscow. Hotline-Telecom. 2014. 210 p. (In Russ.).
8. Bosi M., Goldberg R.E. Introduction to digital audio coding and standards. Springer Science+Business Media. USA. 2003. 434 p.
9. Wickerhauser M.V., Adaptive Wavelet Analysis from Theory to Software. Massachusetts: A.K. Peters Ltd. 1994. 486 p.
10. Johnston J.D. Transform coding of audio signals using perceptual noise criteria. *IEEE Transaction on Selected Areas of Communication*. 1988. vol. 6. pp. 314–323.
11. Kovalgin J.A., Vologdin E.I. *Audiotechnika* [Audio techniques]. Moscow. Hotline-Telocom. 2013. 742 p. (In Russ.).
12. Reyes N.R., Candeas P.V. Adaptive signal modeling based on sparse approximations for scalable parametric audio coding. *IEEE Transactions on audio, speech, and language processing*. 2010. vol. 18. pp. 447–460.
13. Petrovsky Al., Herasimovich V., Petrovsky A. Scalable parametric audio coder using sparse approximation with frame-to-frame perceptually optimized wavelet packet based dictionary. 138th AES Convention. 2015. paper 9264. 10 p.
14. Mallat S., Zhang Z. Matching pursuits with time-frequency dictionaries. *IEEE Transaction on Signal Processing*. 1993. vol. 41. no. 12. pp. 3397–3415.
15. Chardon G., Necciari T., Balazs P. Perceptual matching pursuit with Gabor dictionaries and time-frequency masking. Proceedings of IEEE ICASSP'2014. 2014. pp. 3126–3130.
16. Ravelli E., Gaeul R., Daudet L. Matching pursuit in adaptive dictionaries for scalable audio coding. Proceedings of EUSIPCO'2008. 2008. pp. 1–5.
17. Petrovsky Al., Azarov E., Petrovsky A. Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding. Signal Processing, Special issue "Fourier Related Transforms for Non-Stationary Signals". 2011. vol. 91. Issue 6. pp. 1489–1504.
18. Petrovsky Al., Herasimovich M., Petrovsky A. Bio-inspired sparse representation of speech and audio using psychoacoustic adaptive matching pursuit. Proceedings of 18th International Conference of SPECOM 2016. 2016. pp. 156–164.
19. Petrovsky Al., Herasimovich V., Petrovsky A. Audio/speech coding using frame-based psychoacoustic optimized time-frequency dictionaries and its performance evaluation. IEEE conference proceedings "Signal processing: algorithms, architectures, arrangements, and applications" (SPA-2016). 2016. pp. 225–229.
20. Burrus C.S., Gopinath R.A., Guo H. Introduction to wavelets and wavelet transforms. N.J.: Prentice Hall. 1998. 298 p.
21. Cohen I., Raz S., Malah D. Orthonormal shift-invariant adaptive packet decomposition and representation. Signal Processing. 1997. vol. 57. Issue 3. pp. 251–270.
22. *Analizatory rechevyh i zvukovyh signalov: metody, algoritmy i praktika / pod red. prof. A.A. Petrovskogo* [Analyzers of speech and audio signals: methods, algorithms and practices. Edited by A.A. Petrovsky]. Minsk: Bestprint. 2009. 455 p.
23. Zwicker E., Fastl H. Psychoacoustics: Facts and Models. Berlin, Germany: Springer-Verlag. 1990. 380 p.
24. ITU-R Recommendation BS.1387, Method for Objective Measurements of Perceived Audio Quality. 1998.
25. Petrovsky Al. A multiresolution auditory model using adaptive WP excitation scalograms. *Polska akademia nauk "Elektronika"*. 2008. vol. 49. no 4. pp. 65–70.
26. Petrovsky Al., Krahe D., Petrovsky A. Real-time wavelet packet-based low bit rate audio coding on a dynamic reconfigurable system. Proc. 138th Convention. 2003. paper 5778. 22p.
27. Petrovsky Al., Rodionov M., Petrovsky A. Dynamic reconfigurable on the lifting steps wavelet packet processor with frame-based psychoacoustic optimized time-frequency

tiling for real-time audio applications. *Design and architectures for digital signal processing*. InTech. 2013. pp. 3–30.
28. Karmakar A., Kumar A., Patney R.K. Synthesis of an optimal wavelet based on auditory perception criterion. *EURASIP Journal on Advance in Signal Processing*. 2011. vol. 2011. no. 1. pp. 170927.
29. Petrovsky Al. [A building the psychoacoustic model in the wavelet domain for the perceptual processing of speech and sound signals]. *Nauchno-prakticheskij zhurnal «Rechevye tehnologii» – Scientific journal "Speech Technologies"*. 2008. vol. 4. pp. 61–71. (In Russ.).
30. Coifman R., Wickerhauser M.V. Entropy-Based Algorithms for Best Basis Selection. *IEEE Transaction on Information Theory*. 1992. vol. 38. no. 2. pp. 713–718.
31. Vera-Candeas P., Ruiz-Reyes N., Roza-Zurera M. Transient modelling by Matching-Pursuits with a wavelet dictionary for parametric audio coding. *IEEE Signal Processing Letters*. 2004. vol. 11. no. 3. pp. 349–352.
32. Petrovsky Al., Petrovsky A. Matching pursuit algorithm with frame-based auditory optimized WP-dictionary for audio transient modeling. *Polska academia nauk "Elektronika"*. 2008. vol. 49. no. 4. pp. 74–79.
33. Heusdens R., Vafin R., Kleijn W.B. Sinusoidal modeling using psychoacoustic-adaptive matching pursuits. *IEEE Signal Processing Letters*. 2002. vol. 9. no. 8. pp. 262–265.
34. Huber R., Kollmeier B. PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on audio, speech, and language processing*. 2006. vol. 14. pp. 1902–1911.
35. Vos K., Sørensen K. V., Jensen S. S., Valin J.-M., Voice coding with Opus. Proc. AES 135th Convention. 2013. paper 8941.10 p.
36. Valin J.-M., Maxwell G., Terriberry T.B., Vos K. High-quality, low-delay music coding in the Opus codec. Proc. AES 135th Convention. 2013. paper 8942. 10 p.