УДК [621.391 + 517.443] : 004.934

Анализ алгоритма кодирования аудио волны на основе спектрограмм

А.А. Жарких[1], И.А. Павлов[2]

[1] *Судоводительский факультет МА МГТУ, кафедра радиотехники и радиотелекоммуникационных систем*

[2] *Политехнический факультет МГТУ, кафедра высшей математики и программного обеспечения электронно-вычислительных машин*

Аннотация. Рассмотрен алгоритм кодирования аудио волны (АКАВ), а также алгоритм восстановления аудио волны после АКАВ для её хранения в стандартных форматах и воспроизведения. Описывается множество признаков, формируемых на основе АКАВ. Это множество содержит вектор модулей ординат глобальных экстремумов и вектор разностей абсцисс соседних глобальных экстремумов. Анализируется возможность использования данного множества в системе распознавания аудио сигналов. На основе различных показателей проводится сравнение исходного аудио сигнала и восстановленного после АКАВ.

Abstract. In this work the algorithm of audio wave coding (AAWC) and the algorithm of audio wave recovery after AAWC for storing this wave and its playback have been considered. The set of features extracted on the base of AAWC have been described. This set contains the vector of constant-sign intervals and the vector of global extremes on each of these intervals. The possibility of using this set in the audio signals recognition system has been analyzed. On the base of various factors the comparison of source audio signal and signal recovered after AAWC has been realized.

Ключевые слова: распознавание аудио сигналов, распознавание речи, распознавание изолированных слов, информативные признаки, анализ во временной области, алгоритм кодирования аудио волны, спектральный анализ, дискретное преобразование Фурье, спектрограмма

Key words: audio signals recognition, speech recognition, isolated words recognition, informative features, time-domain analysis, algorithm of audio wave coding, frequency-domain analysis, discrete Fourier Transform, spectrogram

1. Введение

Цель работы – количественная и визуальная оценка изменений в аудио сигнале после использования алгоритма кодирования аудио волны (АКАВ).

В работах (*Лейтес, Соболев*, 1969; *Соболев*, 2006) был предложен алгоритм кодирования речевой волны (АКРВ). Авторы алгоритма утверждали, что восстановленный после кодирования речевой сигнал имеет приемлемую разборчивость при прослушивании. Мы использовали данный алгоритм в системе распознавания изолированных слов русского языка для формирования признаков. Тестирование различных вариантов алгоритма показало изменение разборчивости анализируемого сигнала в широком диапазоне. Результаты распознавания кодированных фрагментов давали также различную точность распознавания. Эти результаты потребовали от нас более тщательного математического анализа АКРВ. В силу того, что мы стали применять этот алгоритм к различным аудио сигналам, мы перешли от авторского названия алгоритма кодирования речевой волны к АКАВ.

В данной работе коротко излагаются алгоритмы кодирования аудио сигнала и обратного восстановления на основе АКАВ. После этого описывается алгоритм распознавания (*Жарких, Павлов*, 2008; *Павлов, Жарких*, 2007), основанный на параметрах кода аудио волны. Далее приведены результаты сравнения исходных аудио сигналов с аудио сигналами, преобразованными алгоритмами кодирования и восстановления на основе АКАВ. Сравнение проводится во временной, частотной и частотно-временной областях.

2. Формирование информативных признаков на основе АКАВ

Под признаком понимается некий параметр исходного сигнала, отражающий свойство, важное для распознавания. Выделять информативные признаки аудио сигнала можно как во временной, так и в частотной области. Для получения признаков, описывающих аудио волну, применялся алгоритм кодирования аудио волны (АКАВ), использующий временное представление аудио сигнала. АКАВ осуществляет поиск глобальных экстремумов на интервалах постоянного знака аудио волны. Исходной информацией для алгоритма является массив дискретных значений аудио сигнала $x = (x_0, x_1, \ldots, x_n, \ldots, x_{L-1})$ и количество отсчетов L в этом массиве. На выходе алгоритм формирует два результирующих вектора:

вектор модулей ординат глобальных экстремумов $y = (y_1, y_2,\ldots, y_j, \ldots, y_J)$, где $y_j = \max |x_n|$ на j-ом интервале постоянного знака аудио волны; вектор разностей абсцисс соседних глобальных экстремумов $t = (t_1, t_2,\ldots, t_j, \ldots, t_J)$, где $t_j = \arg y_j - \arg y_{j-1}$ (величины t_j выражаются в количестве шагов дискретизации кодируемого аудио сигнала). Совокупность двух указанных векторов является компактным описанием аудио волны, которая может быть восстановлена по правилу (*Соболев*, 2006):

$$\hat{x}_n = \frac{(-1)^{j-1}\cdot y_{j-1} + (-1)^j \cdot y_j}{2} + (-1)^{j-1}\cdot \frac{y_{j-1}+y_j}{2}\cdot \cos(\frac{\pi}{t_j}\cdot i), \qquad (1)$$

где $i = 1..t_j$, $j = 1..J$. Таким образом, для каждого аудио сигнала получается вектор информативных признаков: $(y_1, y_2,\ldots, y_J, t_1, t_2,\ldots, t_J)$, состоящий из $2J$ компонент. Эти признаки в дальнейшем используются при распознавании сигналов. АКАВ применялся совместно с низкочастотной Фурье-фильтрацией (*Гольденберг и др.*, 1990), что позволило гибко управлять размером вектора информативных признаков.

3. Алгоритм распознавания аудио сигнала на основе АКАВ признаков

Для распознавания аудио сигналов использовался метод сравнения с эталонами с последующим нахождением степени сходства с эталонами. Степень сходства между аудио записями и эталонами рассчитывалась на основе алгоритма динамического программирования (*Рабинер, Шафер*, 1981).

На вход алгоритма подавались входной и эталонный векторы информативных признаков: $(y_1, y_2,\ldots, y_i,\ldots, y_M, t_1, t_2,\ldots, t_i,\ldots, t_M)$, $(Y_1, Y_2,\ldots, Y_j,\ldots, Y_N, T_1, T_2,\ldots, T_j,\ldots, T_N)$. Алгоритм дает возможность найти функции f_y и f_Y, позволяющие для любого элемента входного вектора признаков найти соответствующий ему элемент эталонного вектора признаков. На основе данного алгоритма определялась степень сходства входного и эталонного векторов признаков.

Степень сходства между парами (y_i, t_i) и (Y_j, T_j) рассчитывалась по формуле:

$$R_{i,j} = \left(\frac{\min\{y_i, Y_j\}}{\max\{y_i, Y_j\}}\cdot \omega_1 + \frac{\min\{t_i, T_j\}}{\max\{t_i, T_j\}}\cdot \omega_2 \right) / (\omega_1 + \omega_2), \qquad (2)$$

где $i = 1,\ldots, M$; $j = 1,\ldots, N$; ω_1, ω_2 – весовые коэффициенты, $\omega_1 + \omega_2 = 1$.

Алгоритм распознавания показал различную степень правильного распознавания изолированных слов русского текста. Если использовались дополнительные фильтры, то степень распознавания изменялась от 50 до 97 процентов. При кодировании АКАВ разборчивость аудио сигнала как правило ухудшалась. Однако прямой корреляции между ухудшением качества распознавания и ухудшением разборчивости при прослушивании не наблюдалось. То есть были варианты приемлемые при прослушивании и хорошие по распознаванию, но были и варианты плохие при прослушивании и хорошие при распознавании. Это и привело авторов к необходимости тщательного математического анализа результатов применения АКАВ.

4. Сравнение исходного сигнала и восстановленного после АКАВ

Для различных вариантов аудио сигналов были проведены сравнения исходных записей с записями, восстановленными после АКАВ. Сравнения проводились во временной, частотной и частотно-временной областях. Рассматривались три варианта образцов: фрагменты записей речевых сигналов фиксированного говорящего, фрагменты записей классической музыки, фрагменты записей современной музыки.

Для анализа поведения аудио сигнала часто используют спектрограммы. Спектрограмма представляет собой функцию двух переменных: время и частота. Таким образом, аудио сигнал, представляющий собой функцию времени, т.е. функцию одной переменной, преобразуется в спектрограмму являющейся функцией двух переменных.

Для визуализации речи использовалась спектрограмма на основе кратковременного преобразования Фурье. Термин кратковременное преобразование Фурье означает, что преобразование Фурье осуществляется на коротких временных участках аудио сигнала по сравнению с его длительностью. Традиционно в таком преобразовании используется специальное окно, позволяющее определенным образом сгладить данные.

Для построения спектрограммы аудио сигнал разбивается на короткие сегменты одинаковой длительности по времени. К каждому из этих сегментов применяется быстрое преобразование Фурье (естественно, аудио сигнал записан в виде дискретных отсчетов). На каждом из сегментов спектр является комплексно-значной функцией номера отсчета (или момента времени). Известно, что комплексно-значную функцию невозможно построить в одной системе координат на плоскости. Поэтому традиционно при анализе спектра строят амплитудный и фазовый спектр любого сигнала. Амплитудный спектр представляет собой модуль комплексного спектра, а фазовый – его аргумент. Спектрограмма

представляет собой объединение амплитудных спектров, вычисленных на коротких сегментах, в функцию двух переменных или матрицу.

Согласно многим источникам и самостоятельным экспериментам авторов, амплитудный спектр плохо представляется в линейном масштабе. Это плохое представление характерно как для кратковременного преобразования Фурье, так и для преобразования Фурье на более длинных реализациях. Это связано с тремя обстоятельствами: с особенностью человеческого зрения (ограниченная разрешающая способность и нелинейное восприятие изображения), с особенностями представления изображений и графиков на компьютере, а также с конкретными значениями амплитудных спектров, возникающих в процессе вычисления. Если преобразование Фурье осуществляется на длинных отрезках записи, то типичной является ситуация, когда несколько низкочастотных составляющих имеют очень большое значение, а большое количество (70-85 %) более высокочастотных составляющих имеют существенно меньшие значения и представляются как будто бы шумом. При этом отличие между максимальным значением низкочастотной составляющей и более высокочастотной составляющей может составлять несколько десятков порядков. Если преобразование Фурье осуществляется на коротких промежутках, то возможна ситуация, когда амплитудный спектр имеет некий шумоподобный характер. Выделить вклад определенных гармоник оказывается очень сложно. В частности, при сборке кратковременных амплитудных спектров в спектрограмму проявляется только часть спектральных составляющих, а б***о***льшая часть теряется.

Традиционно для визуализации амплитудного спектра аудио сигнала и спектрограммы, в частности, используется логарифмическая шкала в децибелах:

$$S(k) = 20 \cdot \log_{10} |X(k)|, \tag{3}$$

где $k = 0,\ldots,N\text{-}1$, N – количество отсчетов в спектре, $|X(k)|$ – k-е значение отчета амплитудного спектра исходного сигнала, $S(k)$ – результирующее значение отчета амплитудного спектра. Однако эта шкала, на взгляд авторов, является неудобной, т.к. в ней неправильно обрабатываются нулевые значения амплитуд. Если значение амплитуды становится близким к нулю, но положительным, то в логарифмической шкале это соответствует большому отрицательному значению. Минимальное значение амплитуды, которое можно представить в шкале децибел, не нарушив физического смысла, равно 1.

Авторы используют альтернативный способ визуализации амплитудного спектра и спектрограммы, в частности, в работах (*Жарких, Коннов*, 2007; *Zharkikh, Pavlov*, 2008). Визуализация спектрограммы проводится на основе функции гиперболического тангенса и представлена формулой:

$$A(r,k) = \left[255 \cdot th(\alpha \cdot |X_r(k)|)\right], \tag{4}$$

где $|X_r(k)|$ – k-е значение отсчета амплитудного спектра сегмента r, α – параметр для управления визуализацией, $A(r, k)$ – значение пикселя изображения спектрограммы, хранящейся в виде матрицы, в которой r – индекс столбца, соответствующий диапазону временной шкалы $t_r = 0, RT, 2RT,\ldots, (N_R\text{-}1)RT$ спектрограммы и k – индекс строки, соответствующий диапазону частотной шкалы $F_k = kF_s/N$, $k = 0, 1, \ldots, N/2$ спектрограммы, T – период дискретизации сигнала, F_s – частота дискретизации сигнала. В выражении (4) кратковременное преобразование Фурье вычисляется согласно формуле (*Rabiner, Schafer*, 2009):

$$X_r(k) = \sum_{m=rR}^{rR+L-1} x(m)w(rR-m)e^{-j\frac{2\pi}{N}km}, \tag{5}$$

где L – размер сегмента (в отсчетах), N – количество дискретных отсчетов, используемых для вычисления быстрого преобразования Фурье (БПФ), $w(m)$ – окно, используемое для вычисления кратковременного преобразования Фурье, N_R – количество сегментов, на которые разбивается сигнал, R – смещение сегмента (в отсчетах).

Множитель 255 выбран из тех соображений, чтобы все значения амплитудного спектра были представлены на картинке в градациях серого. К сожалению, авторы пока не разработали методику оптимального подбора параметра α и подбирают его в процессе вычисления.

Функция гиперболического тангенса преобразует интервал [0; +∞) в интервал [0; 1), поэтому любые значения амплитуды будут отображены на рисунке спектрограммы. Кроме этого, рисунок является более качественным и контрастным, чем это позволяет сделать шкала децибел. Авторы предполагают и дальше разрабатывать эту методику визуализации, т.к. результаты (*Жарких, Коннов*, 2007; *Zharkikh, Pavlov*, 2008) и данной работы показывают, что такая визуализация позволяет выявить некоторые детали и особенности спектра, что не всегда позволяют сделать другие средства.

При моделировании использовались следующие значения параметров аудио сигналов и параметров вычисления спектрограммы:

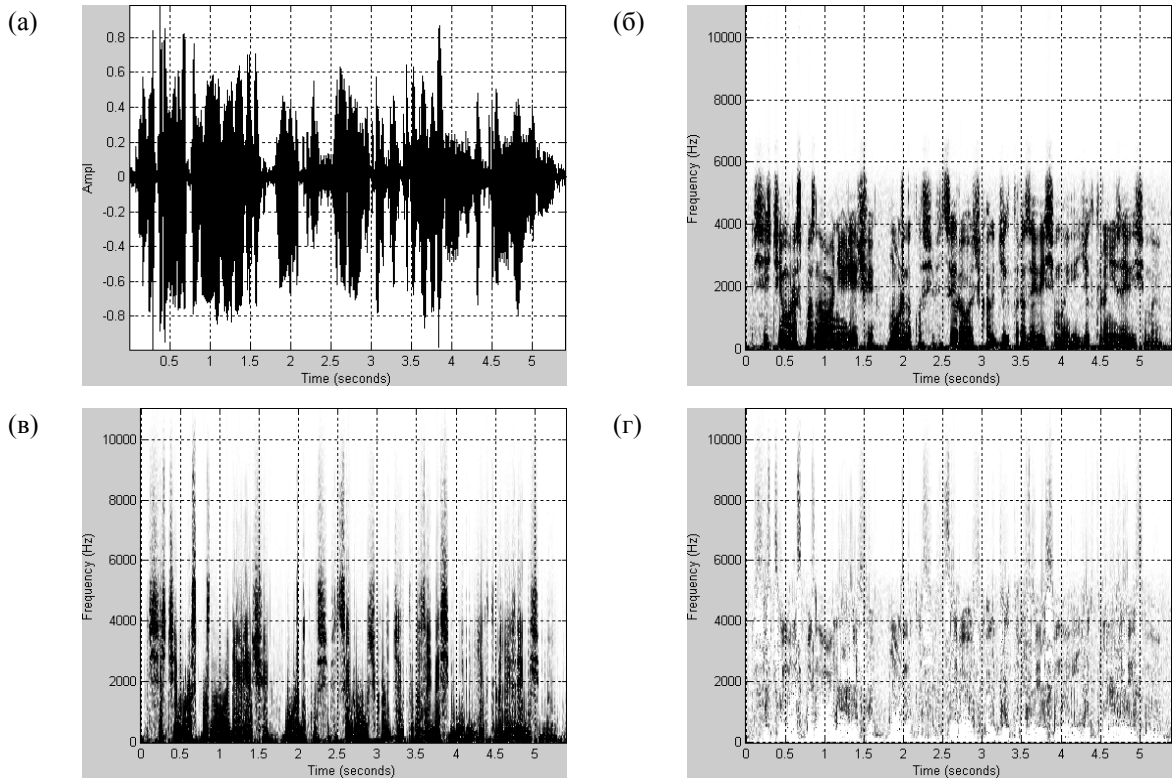

Рис. 1. Фрагмент речевого сигнала, соответствующий фразе, произнесенной одним из авторов

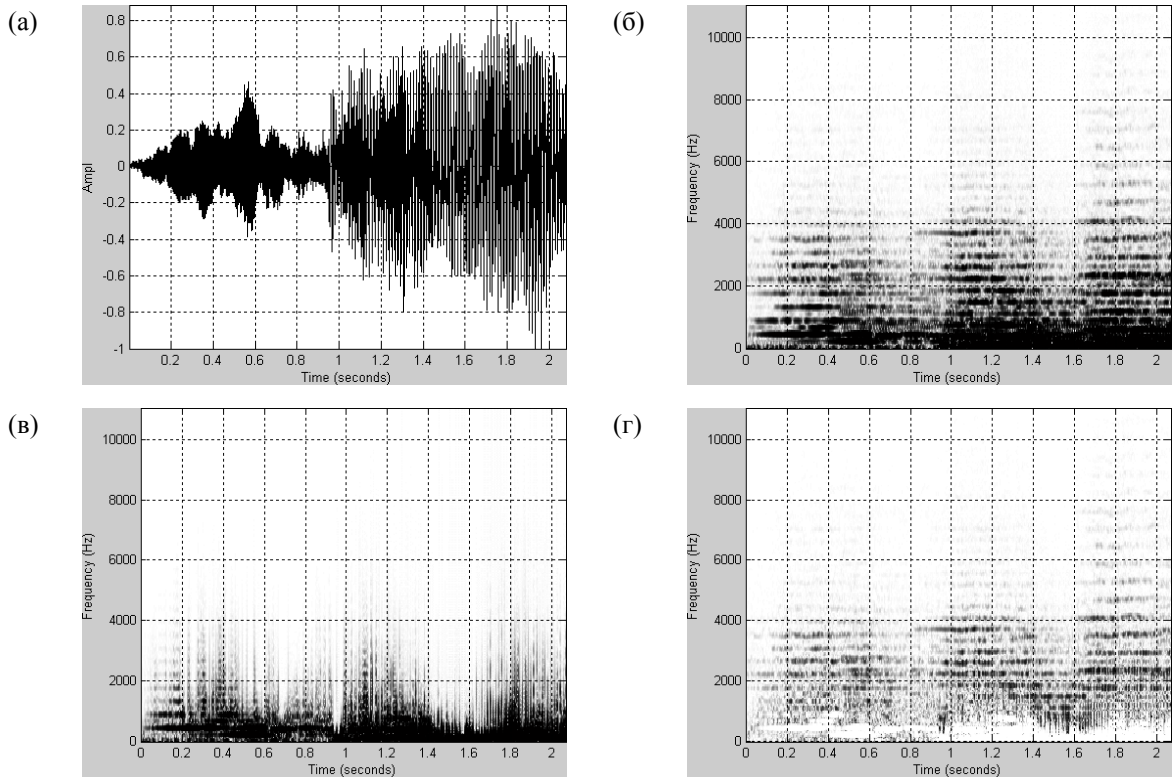

Рис. 2. Фрагмент классической музыки, соответствующий музыкальному произведению "Менуэт", композитор Вольфганг Амадей Моцарт

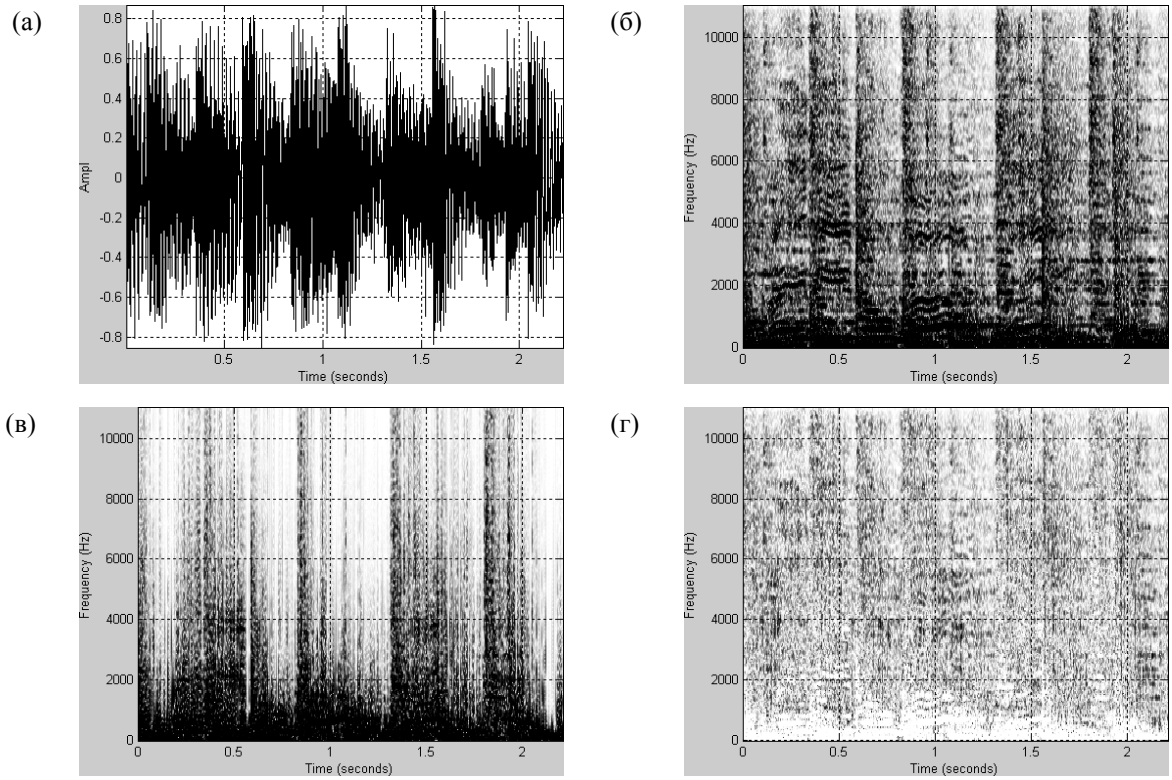

Рис. 3. Фрагмент современной музыки, соответствующий песне "What Is Love" музыканта Haddaway

- Формат аудио сигнала: PCM WAV;
- Частота дискретизации (F_s): 22050 Гц;
- Число уровней квантования (разрядность): 16 бит;
- Окно ($w(n)$): окно Хэмминга;
- Размер сегмента (L): 512 отсчетов (23 мс);
- Смещение сегмента (R): 170 отсчетов (8 мс);
- Перекрытие сегментов (L-R): 342 отсчета (15 мс);
- Размер БПФ (N): 512 отсчетов (23 мс);
- Параметр для управления визуализацией (α): 1.

Для удобства сравнения исходного сигнала и восстановленного после АКАВ строилось изображение разности спектрограмм этих сигналов. Для амплитуд спектрограмм $A_1(r,\ k)$ и $A_2(r,\ k)$ изображение разности строилось с использованием формулы:

$$A_3(r,k) = \left| A_1(r,k) - A_2(r,k) \right|. \qquad (6)$$

Несколько характерных примеров спектрограмм приведены на рис. 1-3. На всех рисунках: график (а) – исходный сигнал, график (б) – спектрограмма исходного сигнала, график (в) – спектрограмма сигнала, восстановленного после АКАВ, график (г) – разность спектрограмм исходного сигнала и восстановленного после АКАВ.

Кроме этого проводились следующие оценки, которые осуществлялись на основе метрики L_2:

- Нормированное расстояние между исходным и восстановленным после АКАВ сигналом:

$$\rho(x,y) = \|x-y\| / (\|x\| + \|y\|), \qquad (7)$$

в выражении (5)

$$\|x\| = \sqrt{\sum_{m=0}^{N-1} x_m^{\,2}},$$

где N – количество временных отсчетов, x_m – значение отчета исходного сигнала, y_m – значение отчета сигнала, восстановленного после АКАВ. Аналогичным образом рассчитывались $\|y\|$ и $\|x\text{-}y\|$.

- Коэффициент корреляции во временной области между исходным и восстановленным после АКАВ сигналом:

$$k(x,y) = (x,y)/(\|x\| \cdot \|y\|), \tag{8}$$

в выражении (6)

$$(x,y) = \sum_{m=0}^{N-1} x_m \cdot y_m \ .$$

- Коэффициент корреляции в частотной области между исходным и восстановленным после АКАВ сигналом:

$$K(X,Y) = \mathrm{Re}(\sum_{m=0}^{N-1} X_m \cdot \overline{Y_m})/(\|X\| \cdot \|Y\|) \ . \tag{9}$$

5. Заключение

Результаты анализа АКАВ позволяют сделать следующие выводы:

1) Сигнал, полученный в результате кодирования на основе АКАВ, требует для хранения объем памяти в 4-5 раз меньше, чем исходный сигнал.
2) Во всех случаях действие АКАВ эквивалентно пропусканию сигнала через фильтр нижних частот.
3) Во многих случаях применение АКАВ приводит также к режекции средней части спектра в области нижних частот.
4) Нормированное расстояние между исходным и восстановленным после АКАВ сигналом для аудио сигналов различного класса составляет приблизительно 0.22-0.5.
5) Коэффициент корреляции во временной области между исходным и восстановленным после АКАВ сигналом для различных типов аудио сигналов изменяется от 0.5 до 0.92.
6) Коэффициент корреляции в частотной области между исходным и восстановленным после АКАВ сигналом для различных типов аудио сигналов изменяется от -0.24 до 0.35. Такие маленькие величины связаны с изменением фазы в восстановленном сигнале и интерференцией сигналов при вычислении коэффициента корреляции.
7) Анализируя графики разности спектрограмм исходного сигнала и восстановленного после АКАВ, можно сделать вывод, что для всех трех вариантов образцов лучше всего сохраняется частотный диапазон 0-1000 Гц.

Литература

Rabiner L.R., Schafer R.W. Theory and application of digital speech processing. *Prentice Hall Inc.*, 2009. (In preparation)

Zharkikh A., Pavlov I. Audio signal feature extraction based on the algorithm of audio wave coding. *Pattern Recognition and Image Analysis: New Information Technologies: Conference Proceedings, Nizhny Novgorod*, v.2, p.355-358, 2008.

Гольденберг Л.М., Матюшкин Б.Д., Поляк М.Н. Цифровая обработка сигналов. *М., Радио и связь*, 256 с., 1990.

Жарких А.А., Коннов Е.В. Управляемая визуализация спектра изображения. *Докл. Всеросс. конф. "Математические методы распознавания образов - 13", М., МАКС Пресс*, с.319-323, 2007.

Жарких А.А., Павлов И.А. Реализация программного модуля распознавания речевых сигналов. *Сборник материалов VIII Междунар. конференции "Распознавание-2008", Курск, Курск. гос. техн. ун-т*, ч. 1, с.158-159, 2008.

Лейтес Р.Д., Соболев В.Н. Цифровое моделирование систем синтетической телефонии. *М., Связь*, 120 с., 1969.

Павлов И.А., Жарких А.А. Программный модуль выделения информативных признаков речевого сигнала. *Материалы 15 межрегиональной научно-техн. конференции "Обработка сигналов в системах наземной связи и оповещения", М., НТОРЭС им. А.С. Попова*, с.223-224, 2007.

Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. *М., Радио и связь*, 496 с., 1981.

Соболев В.Н. Простые алгоритмы экономного кодирования и декодирования речевой волны. *Материалы 14 межрегиональной научно-техн. конференции "Обработка сигналов в системах наземной связи и оповещения", М., НТОРЭС им. А.С. Попова*, с.172-174, 2006.