Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Факультет компьютерного проектирования

Кафедра электронно-вычислительных средств

А. А. Петровский, М. И. Вашкевич, И. С. Азаров

ЦИФРОВАЯ ОБРАБОТКА АУДИО- И ВИДЕОДАННЫХ

Рекомендовано УМО по образованию в области информатики и радиоэлектроники в качестве пособия для специальности 1-40 80 01 «Элементы и устройства вычислительной техники и систем управления»

Минск БГУИР 2017

УДК 621.391(076)
ББК 32.811.3я73
П30

Рецензенты:
кафедра информационных систем и технологий
Белорусского национального технического университета
(протокол №1 от 07.09.2015);

доцент кафедры информатики филиала «Минский радиотехнический колледж» учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»,
кандидат технических наук, доцент В. Г. Лукьянец

Петровский, А. А.

П30 Цифровая обработка аудио- и видеоданных : пособие / А. А. Петровский, М. И. Вашкевич, И. С. Азаров. – Минск : БГУИР, 2017. – 64 с. : ил.
ISBN 978-985-543-269-3.

Пособие посвящено изложению проблемы подавления шума в сигналах, содержащих речевые сообщения. Дано описание классического подхода к решению задачи шумоподавления, основанного на методе вычитания спектров. Рассмотрены вопросы улучшения качества восстановленного речевого сигнала путем применения техники анализа на основе дискретного преобразования Фурье с неравномерным частотным разрешением. Для эффективного подавления реверберационных и полигармонических шумов описан специальный класс алгоритмов, основанный на обработке сигнала в модуляционной области. Приведены примеры описания систем шумоподавления в среде Matlab.

УДК 621.391(076)
ББК 32.811.3я73

ISBN 978-985-543-269-3

Содержание

Введение

Широкое распространение коммуникационных устройств для аудио- и видеосвязи ставит перед разработчиками подобных систем разнообразные задачи, одной из которых является подавление шума во входном аудиопотоке.

Задаче очистки речевых сигналов от аддитивного шума уделяется значительное внимание. Столь устойчивый интерес обусловлен широким кругом возможных применений и ограничениями существующих алгоритмов. Речевые сигналы, зарегистрированные в своей естественной акустической обстановке, могут содержать значительные искажения, обусловленные фоновым шумом, голосами других дикторов и т. д. Наличие аддитивного шума часто затрудняет восприятие речевого сообщения слушателем, а также значительно усложняет процесс автоматической обработки в таких задачах, как распознавание речи, идентификация диктора, модификация речи в слуховых аппаратах, кодирование и т. д. Для повышения работоспособности соответствующих алгоритмов обработки зашумленная речь, как правило, должна быть очищена от посторонних звуков системой автоматического шумоподавления. Поиск наиболее оптимального метода шумоочистки для того или иного вида помех должен выполняться при помощи объективных и субъективных оценок. На выбор алгоритма также влияет интенсивность шума, которая в зависимости от приложения варьируется от экстремальной до умеренной (соотношение сигнал/шум – $10 \leq \mathrm{SNR} \leq 20$ дБ).

Большинство существующих систем шумоподавления работают в частотной области, используя вариации метода спектрального взвешивания. К сожалению, его негативной особенностью является появление в реконструированном речевом сигнале искажений, известных как «музыкальные тона». Было предложено много подходов, чтобы устранить этот феномен, включая подходы, основанные на восприятии человеком звуковой информации. Интересным обобщением методов спектрального взвешивания является обработка зашумленного речевого сигнала в подпространствах. Оценка речи здесь рассматривается как задача оптимизации с ограничениями, где искажения речевого сигнала минимизируются с учетом остаточной мощности шума.

1 СИСТЕМЫ ПОДАВЛЕНИЯ ШУМА В РЕЧЕВЫХ СИГНАЛАХ

1.1 Общие теоретические сведения

Цель обработки речевого сигнала, записанного в условиях акустического загрязнения, состоит в повышении его качества путем уменьшения фонового шума без снижения разборчивости речевого сообщения.

Наиболее распространенным методом подавления шума является спектральное вычитание. Как и в большинстве методов улучшения качества речевого сигнала в методе спектрального вычитания делается предположение, что спектр мощности зашумленного речевого сигнала равен сумме спектра чистого сигнала и спектра некоррелированного шума. Это предположение является обоснованным при анализе спектра на коротких временных интервалах (порядка 25 мс) и ведет к построению простого метода спектрального вычитания.

Базовый метод спектрального вычитания состоит в вычислении спектра мощности для каждого сегмента входного сигнала, умноженного на оконную весовую функцию, и вычитании спектра мощности из полученного спектра зашумленного сигнала. Оценка спектра мощности шума производится по сегментам сигнала, в которых отсутствует речь. Информация о фазе частотных компонент для синтеза сигнала, очищенного от шума, берется из ДПФ сегмента исходного сигнала.

1.2 Спектральное вычитание: базовый алгоритм

Допустим, что $x(n)$ – это входной зашумленный сигнал, который состоит из чистого речевого сигнала $s(n)$ и аддитивного шумового сигнала $d(n)$, т. е.

$$x(n) = s(n) + d(n). \tag{1.1}$$

Применяя дискретное преобразование Фурье (ДПФ) к правой и левой части выражения, получаем

$$X(\omega) = S(\omega) + D(\omega). \tag{1.2}$$

Можно выразить $X(\omega)$ в полярной системе координат:

$$X(\omega) = |X(\omega)|e^{j\varphi_x(\omega)}, \tag{1.3}$$

где $|X(\omega)|$ и $\varphi_x(\omega)$ – амплитудный и фазовый спектры зашумленного сигнала соответственно.

Спектр шума $D(\omega)$ тоже можно выразить в терминах амплитудного и фазового спектров как $D(\omega) = |D(\omega)|e^{j\varphi_d(\omega)}$. Амплитудный спектр шума $|D(\omega)|$ в общем случае неизвестен, но может быть заменен своим усредненным значением, вычисленным по фреймам входного сигнала, в которых отсутствует речь.

Подобным образом фазовый спектр шума $\varphi_d(\omega)$ можно заменить фазовым спектром зашумленной речи $\varphi_x(\omega)$. Частично это объясняется тем, что изменение фазы не влияет на разборчивость речи, а может только иметь влияние в некоторой степени на качество речевого сигнала. После приведенных замен можно записать выражение для оценки спектра чистой речи:

$$\hat{S}(\omega) = \left(|X(\omega)| - \left|\widehat{D}(\omega)\right|\right)e^{j\varphi_x(\omega)}, \tag{1.4}$$

где $\left|\widehat{D}(\omega)\right|$ – оценка амплитудного спектра шума, определенная во время пауз речи.

Уравнение (1.4) дает представление о главном принципе, лежащем в основе метода спектрального вычитания. В более общей форме метод спектрального вычитания формулируется следующим образом:

$$\left|\hat{S}(\omega)\right|^p = |X(\omega)|^p - \left|\widehat{D}(\omega)\right|^p, \tag{1.5}$$

где p – показатель степени, при $p = 1$ выражение (1.5) описывает первоначальный метод спектрального вычитания. Часто в практических приложениях используют показатель $p = 2$. В этом случае (1.5) описывает правило вычитания спектров мощности.

Необходимо отметить, что правая часть (1.5) может иметь отрицательный знак, что является следствием неточности в оценке спектра шума. Однако амплитудное значение (или мощность) не может быть отрицательным числом, следовательно необходимо дополнить правило (1.5), для того чтобы оценка спектра чистой речи всегда имела неотрицательные значения. Введем следующие обозначения:

$$P_{\hat{s}}(\omega) = \left|\hat{S}(\omega)\right|^2, \quad P_x(\omega) = |X(\omega)|^2, \quad P_{\hat{d}}(\omega) = \left|\widehat{D}(\omega)\right|^2, \tag{1.6}$$

тогда модифицированное правило для вычитания спектров мощности будет иметь следующий вид:

$$V(\omega) = P_x(\omega) - P_{\hat{d}}(\omega), \tag{1.7}$$

$$P_{\hat{s}}(\omega) = \begin{cases} V(\omega), & \text{если } V(\omega) > 0, \\ 0 & \text{иначе.} \end{cases} \tag{1.8}$$

Очищенный от шума речевой сигнал получается из $P_s'(\omega)$ путем обратного преобразования Фурье:

$$\hat{s}(n) = \text{ОДПФ}\left\{\sqrt{P_{\hat{s}}(\omega)}e^{j\varphi_x(\omega)}\right\}, \quad (1.9)$$

где $\varphi_x(\omega)$ – фазовый спектр исходного фрагмента зашумленного сигнала.

Главная проблема в описанном методе подавления шума состоит в том, что в обработанном сигнале возникает «новый» шум. На слух этот шум воспринимается как музыкальные тона, имеющие хаотический порядок. В литературе этот шум получил название «музыкального шума». Кроме того, несмотря на то что происходит уменьшение шума, его значительная часть остается в обработанном сигнале.

1.3 Природа музыкального шума

Метод спектрального вычитания изначально разрабатывался для очистки сигнала от белого шума. Чтобы объяснить природу музыкального шума, надо учесть, что в спектре белого шума, вычисленного на коротком интервале времени, имеются локальные максимумы и минимумы. Их частотное положение и амплитуда являются случайными и изменяются случайным образом для каждого последующего сегмента. Когда происходит вычитание сглаженной оценки спектра шума из текущего спектра, то локальные максимумы спектра смещаются вниз, а в окрестности минимума устанавливают значение нуля (минус бесконечность на логарифмической шкале). Таким образом, и после операции вычитания в спектре шума остаются локальные максимумы. Наиболее широкие максимумы на слух воспринимаются как изменяющийся широкополосный шум. Более узкие спектральные максимумы, обладающие продолжительностью и смещающиеся по частоте, образуют спектральные «трассы» и воспринимаются как меняющиеся во времени тона, которые и называют музыкальным шумом.

1.4 Спектральное вычитание: модификация алгоритма

Модификация метода спектрального вычитания заключается в минимизации воспринимаемых на слух узкополосных спектральных пиков (максимумов) путем укорачивания их спектральных «трасс». Это достигается путем изменения алгоритма (1.7) – (1.8) следующим образом:

$$V(\omega) = P_x(\omega) - \alpha P_{\hat{d}}(\omega), \quad (1.10)$$

$$P_{\hat{s}}(\omega) = \begin{cases} V(\omega), & \text{если } V(\omega) > \beta P_{\hat{d}}(\omega), \\ \beta P_{\hat{d}}(\omega) & \text{иначе,} \end{cases} \quad (1.11)$$
$$\alpha \geq 1, \quad \text{и } 0 < \beta \ll 1,$$

где β – параметр, определяющий спектральный минимум шума. Схема модифицированного метода показана на рисунке 1.1.

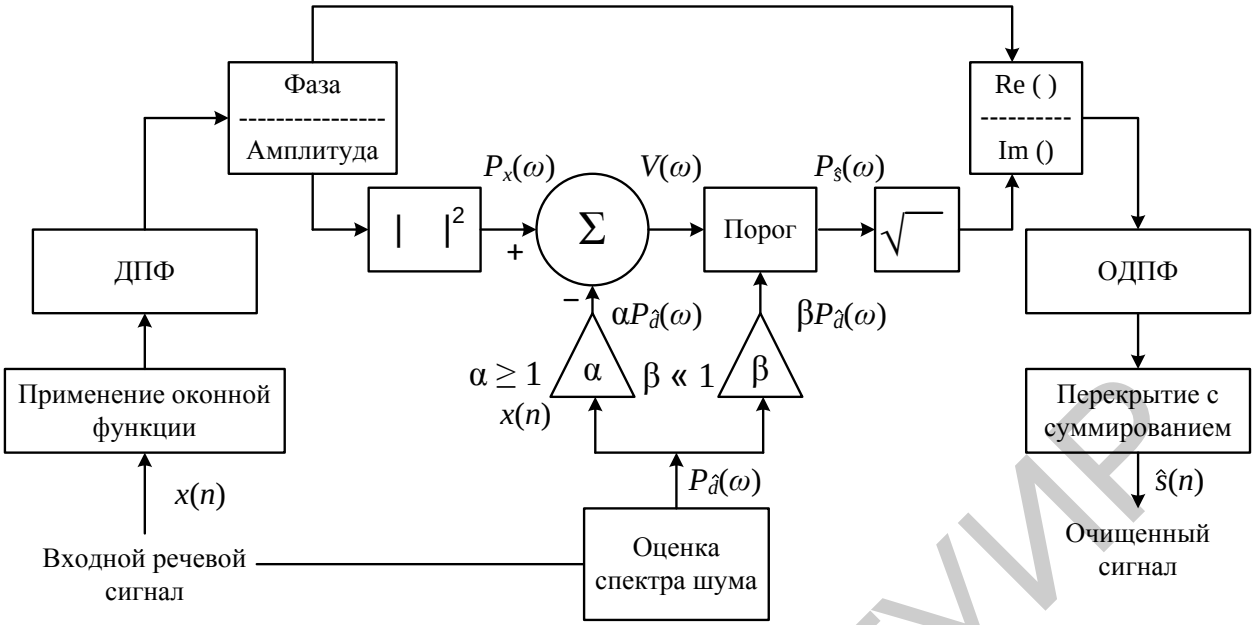


Рисунок 1.1 – Модифицированный метод спектрального вычитания с ограничением спектрального минимума шума

Необходимо отметить, что модифицированный метод (1.10)–(1.11) полностью повторяет (1.7)–(1.8) для значений $\alpha = 1$ и $\beta = 0$.

Из (1.10)–(1.11) можно заключить, что цели уменьшения шумовых спектральных пиков можно достичь, если выбрать $\alpha > 1$, поскольку в этом случае остатки пиков будут ниже относительно случая, когда $\alpha = 1$. Также при $\alpha > 1$ вычитание может удалить практически весь широкополосный шум путем удаления большинства широких пиков. Однако этого недостаточно, поскольку глубокие минимумы, которые находятся вблизи узких пиков, остаются в спектре шума и поэтому «трассы» спектральных пиков остаются длительными. Вторая часть модификации (1.11) заключается в заполнении областей минимума, что делается путем добавления спектрального порога $\beta P_d(\omega)$: не допускается снижение уровня спектральных компонент $P_{\hat{s}}(\omega)$ ниже порогового значения $\beta P_d(\omega)$. Для $\beta > 0$ области минимумов между пиками не столь глубоки, как при $\beta = 0$. Таким образом, спектральные трассы шумовых пиков сокращаются, что уменьшает воспринимаемый музыкальный шум.

1.5 Детектор речевой активности

В простейшем случае детектор речевой активности может быть реализован путем анализа энергии текущего фрейма сигнала и сравнения ее с энергией оценки шума. Фреймы, содержащие речь, в среднем должны обладать большей энергией, чем фреймы, содержащие исключительно шум. На практике чаще всего используется параметр сегментного соотношения сигнал/шум:

$$\text{SegSNR} = 10 \lg \frac{\sum P_x(\omega)}{\sum P_{\hat{d}}(\omega)}. \qquad (1.12)$$

В процессе работы системы шумоподавления для каждого поступающего фрейма рассчитывается значение SegSNR и выполняется сравнение его с пороговым значением N_{thres}, которое определяется экспериментально. Если

$$\text{SegSNR} > N_{thres},$$

то это говорит о том, что текущий сегмент сигнала содержит речевую активность.

На рисункеунок 1.2 показан пример работы детектора речевой активности для реального речевого сигнала, загрязненного белым шумом.

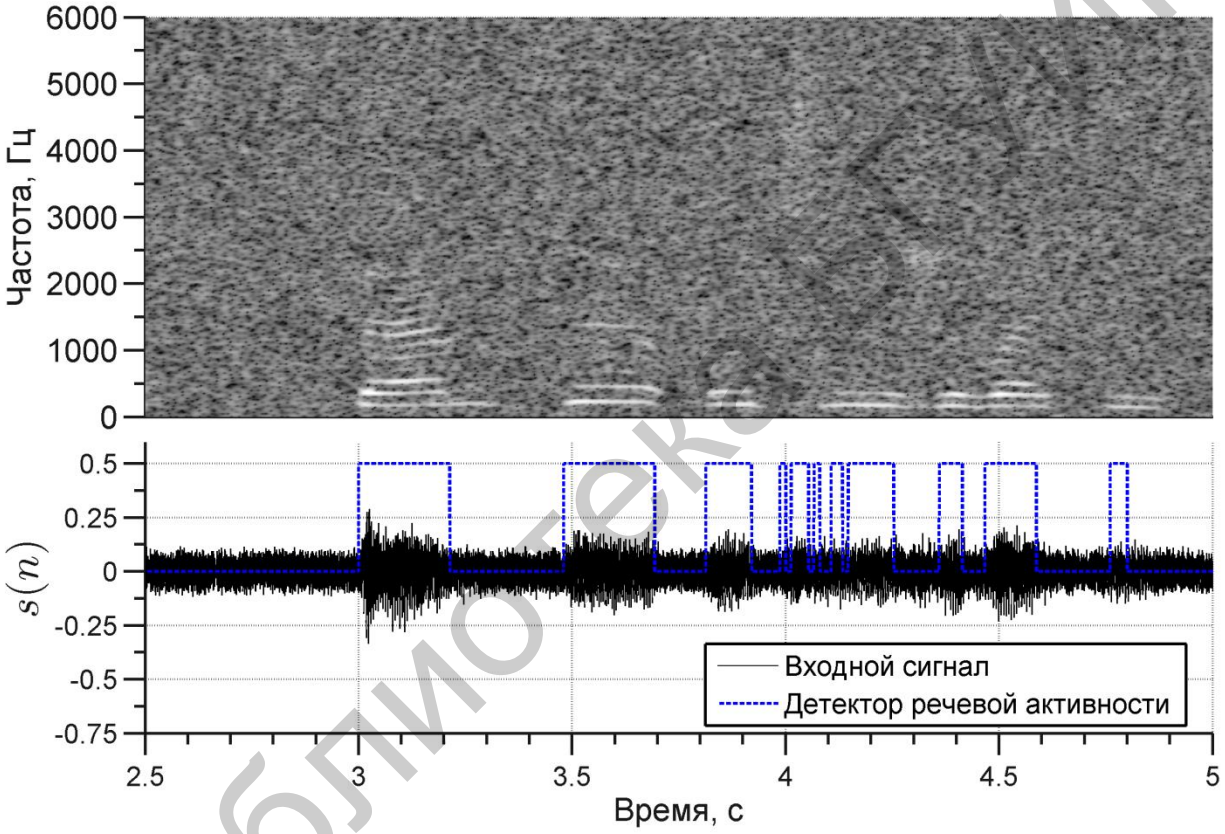


Рисунок 1.2 – Пример работы детектора речевой активности

1.6 Оценка спектра мощности шума

Оценка спектра мощности шума происходит только на тех фреймах входного сигнала, на которых отсутствует речевая активность. При этом оценка производится методом экспоненциального усреднения:

$$P_{\hat{d}}(\omega) = \gamma P_{\hat{d}}(\omega) + (1-\gamma)P_x(\omega), \tag{1.13}$$

где γ – коэффициент усреднения, который обычно выбирается в диапазоне $0{,}9 < \gamma < 1$.

Пример оценки спектра мощности шума показан на рисунке 1.3. Следует отметить, что ошибки в работе детектора речевой активности вызывают появление ложных компонент в спектре мощности шума. Рисунок показывает, что процесс обновления спектра мощности шума прекращается на то время, когда алгоритмом детектируется в сигнале речевая активность.

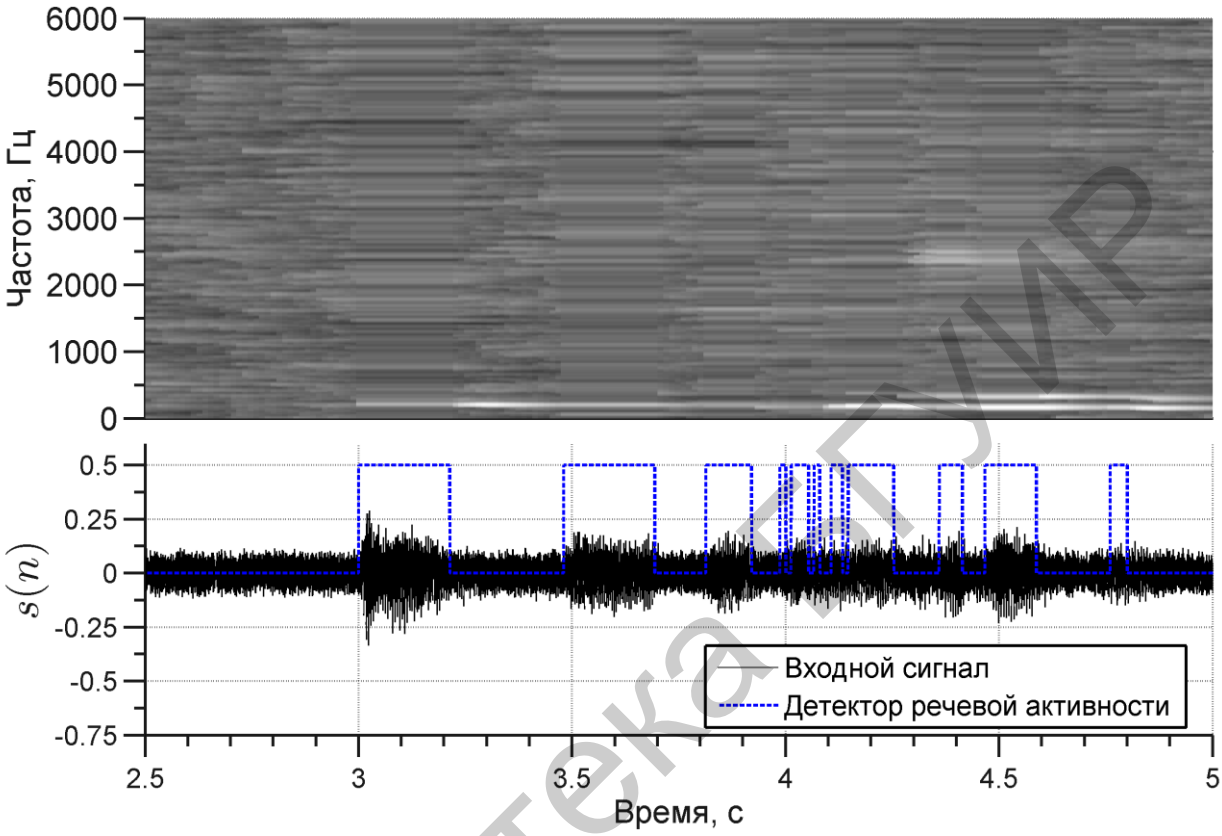


Рисунок 1.3 – Пример выполнения оценки спектра мощности шума

1.7 Реализация базового алгоритма спектрального вычитания в среде Matlab

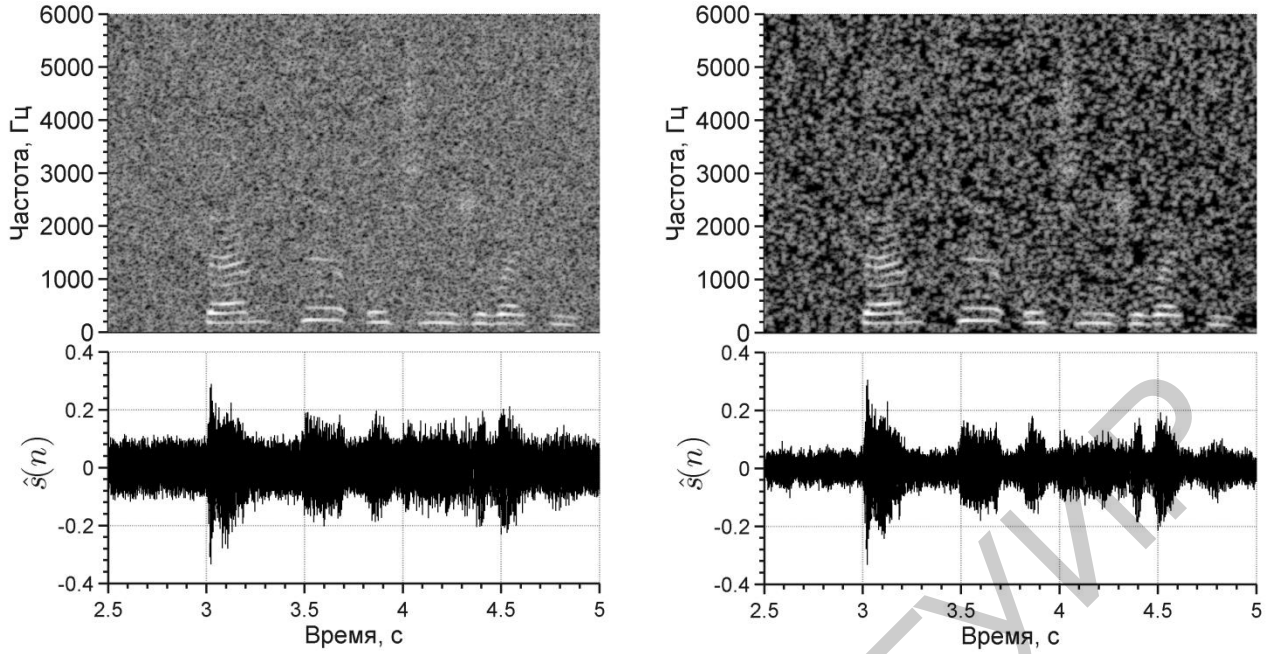
Ниже приводится Matlab-код программы, который реализует базовый алгоритм спектрального вычитания.

```
% Удаление шума методом спектрального вычитания
filename = 'female_high_70_2sec_nos.wav';     % Входной файл
outfile = 'female_high_70_2sec_nos_out.wav'; % Выходной файл

[x, fs] = wavread(filename);     % чтение wav-файла

% === Инициализация переменных ===
frame_duration = 0.02;           % Размер фрейма 0.02 с = 20 мс
Ln = floor(frame_duration*fs);   % Размер фрейма в отсчетах
if rem(Ln,2)==1, Ln=Ln+1; end;   % Размер должен быть четным

% Перекрытие между фреймами (в процентах от размера фрейма)
PERC = 50;
```

```
len1 = floor(Ln*PERC/100);
len2 = Ln-len1;

N_thres = 1.1;  % Пороговое значение ОСШ для детектора речи
beta  = 0.002;  % коэффициент спектрального порога шума
gamma = 0.9;
alpha = 5.01;

win = hamming(Ln); % Окно анализа

% Нормирующий коэффициент для перекрытия с суммированием (50%)
winGain = len2/sum(win); % Нормирующий коэффициент

% Вычисление спектра шума: предполагается, что первые 5 фреймов
% сигнала являются шумом/тишиной

nFFT = 2*2^nextpow2(Ln);     % Длина ДПФ
noise_mean = zeros(nFFT,1); % Буфер для хранения спектра шума
n=1;                         % Индекс входного сигнала
for frame_ind=1:5
   noise_mean=noise_mean+abs(fft(win.*x(n:n+Ln-1),nFFT));
   n=n+Ln;
end
P_d = noise_mean/5;     % Усреднение
P_d = P_d.^2;      % Расчет спектра мощности

% === Выделение памяти и инциализация переменных ===
n=1;                         % Текущий индекс времени
x_old=zeros(len1,1); % Буфер схемы "перекрытие с суммированием"
Nframes = floor(length(x)/len2)-1; % Число фреймов
s_hat = zeros(Nframes*len2,1);     % Буфер для выходного сигнала

%=== Начало обработки ===
%
for frame_ind=1:Nframes

   input_frame = win.*x(n:n+Ln-1);  % Применение оконной функции
   spec = fft(input_frame, nFFT);   % Вычисление ДПФ
   P_x  = abs(spec).^2;   % Вычисление спектра мощности
   theta = angle(spec);    % Вычисление фаз спектральных компонент

   % Сегментного отношения сигнал/шум
   SNRseg=10*log10(sum(P_x)/sum(P_d));

   % Спектральное вычитание
   sub_speech=P_x - alpha*P_d;
   diffw = sub_speech - beta*P_d;

   % Ограничиваем отрицательные компоненты
   z = find(diffw <0);
   if~isempty(z)
```

```
      sub_speech(z) = beta*P_d(z);
   end


   % --- Реализация детектора речевой активности ---
   %
   if (SNRseg < N_thres)     % Обновить оценку спектра шума
      noise_temp = gamma *P_d+(1- gamma)*P_x;
      P_d=noise_temp;  % Новая оценка спектра шума
   end

   % доопределение симметричной части спектра
   sub_speech(nFFT/2+2:nFFT)=flipud(sub_speech(2:nFFT/2));

   x_phase = sqrt(sub_speech).*(cos(theta)+1j*(sin(theta)));

   % Вычисление обратного ДПФ
   xi=real(ifft(x_phase));

   % --- Перекрытие с суммированием ---
   s_hat(n:n+len2-1) = x_old + xi(1:len1);
   x_old = xi(1+len1:Ln);

   n = n + len2;
end

%=== Запись результата ===
wavwrite(winGain*s_hat,fs,16,outfile);
```

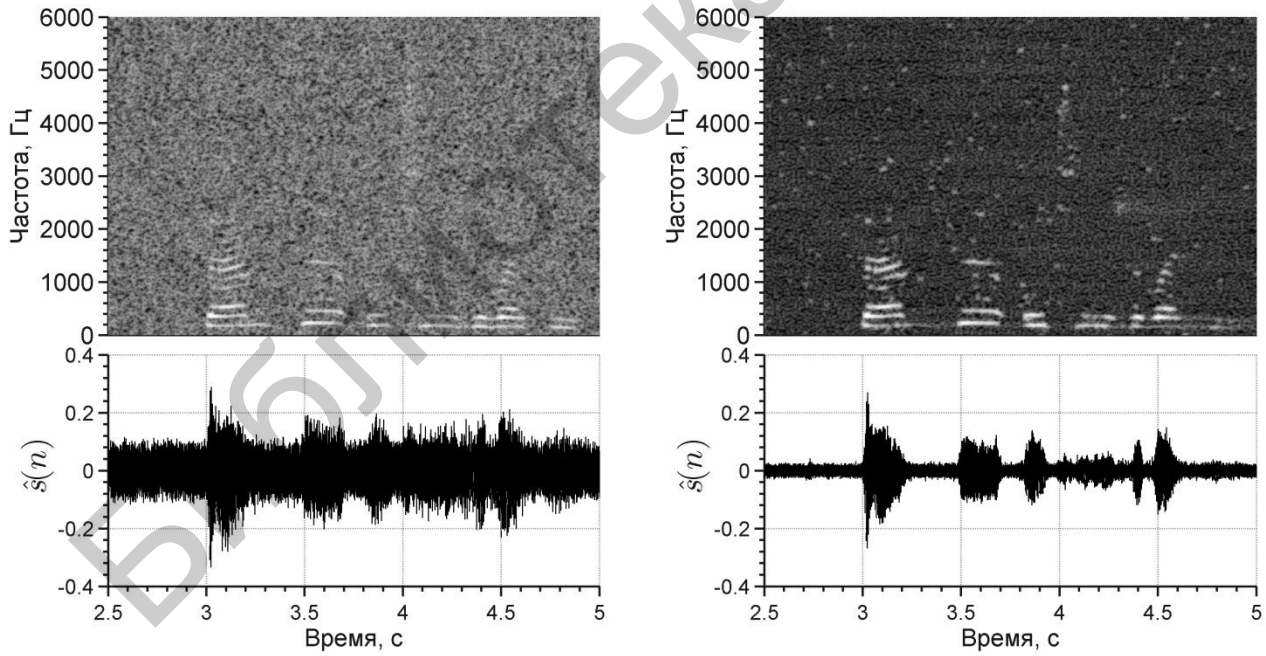
На рисунке 1.4 показан пример работы классического алгоритма вычитания спектра. В спектре выходного сигнала можно видеть значительное присутствие музыкальных шумов. На рисунке 1.5 – модифицированного алгоритма спектрального вычитания с параметрами $\alpha = 5$ и $\beta = 0{,}05$. Выходной сигнал в данном случае имеет более низкий уровень шума.



а б

а – сигнал на входе системы; б – сигнал на выходе системы

Рисунок 1.4 – Работа системы шумоподавления с параметрами $\alpha = 1$ и $\beta = 0$



а б

а – сигнал на входе системы; б – сигнал на выходе системы

Рисунок 1.5 – Работа системы шумоподавления с параметрами $\alpha = 5$ и $\beta = 0{,}05$

2 ОБРАБОТКА РЕЧИ НА ОСНОВЕ ДИСКРЕТНОГО ПРЕОБРАЗОВАНИЯ ФУРЬЕ С НЕРАВНОМЕРНЫМ ЧАСТОТНЫМ РАЗРЕШЕНИЕМ

2.1 Общие теоретические сведения

Дискретное преобразование Фурье (ДПФ) является мощным средством для частотного анализа с постоянной разрешающей способностью. Однако в контексте перцептуальной обработки сигналов, основанной на психоакустической модели восприятия акустической информации человеком, возникает необходимость в неравнополосной частотной декомпозиции сигнала в соответствии со шкалой критических частотных полос (Барков, ERB или MEL). Примером могут служить системы перцептуального кодирования звука и речи, а также подавления шума окружающей среды в речевом сигнале.

Одним из вариантов преобразования с переменным частотным разрешением является дискретное преобразование Фурье (ДПФ) с неравномерным частотным разрешением (от англ. Warped Discrete Fourier Transform (WDFT)), позволяющее получить Z-преобразование конечной последовательности отсчетов сигнала с неравномерным разложением коэффициентов преобразования на единичной окружности Z-плоскости посредством фазового звена.

2.2 Определение WDFT

ДПФ с переменным частотным разрешением (от англ. Nonuniform DFT (NDFT)) является наиболее обобщающим вариантом ДПФ. Кроме единственности решения, NDFT не ограничивает расположение коэффициентов преобразования в Z-плоскости никоим способом. Преобразование WDFT есть специальный случай NDFT, коэффициенты преобразования которого располагаются неравномерно, но регулярно на единичной окружности Z-плоскости.

WDFT последовательности $x[n]$ из N точек определяется по следующей формуле:

$$\hat{X}(z_k) = X(\hat{z}_k) = \sum_{n=0}^{N-1} x[n]\hat{z}_k^{-n}, \quad k = 0,\ldots,N-1, \tag{2.1}$$

где $\hat{z}_k$ – изображения, преобразованных фазовым звеном $A(z)$ равноотстоящих точек на единичной окружности в Z-плоскости:

$$z_k^{-1} = e^{-j\frac{2\pi k}{N}} \quad \rightarrow \quad \hat{z}_k^{-1} = A(z_k) \quad k = 0,\ldots,N-1. \tag{2.2}$$

$A(z)$ – устойчивое фазовое звено произвольного порядка.

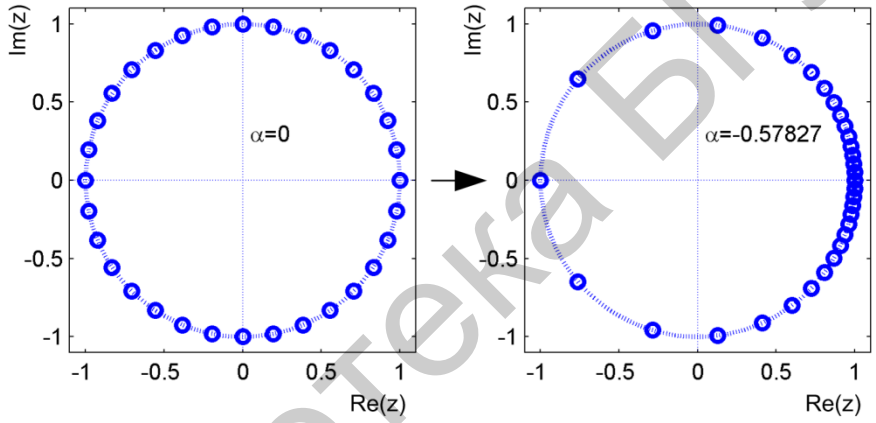
Простейший вариант WDFT основан на фазовом звене первого порядка с действительным коэффициентом a:

$$A(z)=\frac{z^{-1}-a}{1-az^{-1}}. \tag{2.3}$$

Условием стабильности фильтра является $|a|<1$. В зависимости от знака a растягивается низкочастотный $(a>0)$ или высокочастотный $(a<0)$ диапазон спектра путем неравномерного распределения коэффициентов преобразования на единичной окружности в z-плоскости. Формально это может быть выражено следующим образом:

$$\hat{\omega}=\omega+2\,\mathrm{arctg}\left(\frac{a\sin\omega}{1-\cos\omega}\right) \quad \text{для} \begin{cases} z=e^{j\omega}, \\ z=e^{j\hat{\omega}}, \end{cases} \tag{2.4}$$

т. е. осуществляется билинейное преобразование z-плоскости в новую искривленную $\hat{z}$-плоскость (рисунок 2.1).



а б

а – ДПФ; б – WDFT

Рисунок 2.1 – Расположение коэффициентов преобразования на единичной окружности в z-плоскости

Как обобщение ДПФ WDFT также имеет свойства линейности, симметрии и сдвига. Сопряженная симметрия для действительных данных имеет силу и для WDFT:

$$\hat{X}\left(z_{N-1-k}\right)=\hat{X}^{*}\left(z_{k}\right), \tag{2.5}$$

где знак «*» обозначает комплексное сопряжение. Однако ряд важных свойств ДПФ потерян.

В матричной записи (с $\hat{X}[k]$, обозначающим $\hat{X}\left(z_{k}\right)$) WDFT может быть представлено следующим выражением:

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[N-1] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & A(z_0) & \cdots & A(z_0)^{N-1} \\ 1 & A(z_1) & \cdots & A(z_1)^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & A(z_{N-1}) & \cdots & A(z_{N-1})^{N-1} \end{bmatrix}}_{\mathbf{D}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}. \qquad (2.6)$$

2.3 Обратное преобразование WDFT

В случае неравномерного частотного разрешения матрицы Вандермонда, к классу которых относится и матрица **D** WDFT, обладают свойствами, делающими невозможным получение точной обратной матрицы. А именно, матрица WDFT может рассматриваться как сингулярная, так как между некоторыми ее строками существуют почти линейные зависимости. Этот факт представляется как очень малое значение ее детерминанта $\det \mathbf{D} = \prod_{i<j} \left(\hat{z}_i^{-1} - \hat{z}_j^{-1} \right)$ для определенного $\hat{z}_k$. Другими словами, WDFT-матрица является плохо обусловленной. Это означает, что любой численный алгоритм обращения матрицы, применяемый к подобной матрице, является очень чувствительным к малым изменениям данных.

Усиление ошибки может быть оценено посредством использования числа обусловленности матрицы полного ранга (собственного значения матрицы):

$$\mathrm{cond}(\mathbf{D}) = \|\mathbf{D}\| \|\mathbf{D}^{-1}\| = \sigma_{\max} / \sigma_{\min}, \qquad (2.7)$$

где $\| \ \|$ означает произвольную норму матрицы (в общем случае евклидову);

$\sigma_{\max}$ и $\sigma_{\min}$ – наибольшее и наименьшее сингулярные числа.

Как показано на рисунке 2.2, число обусловленности преобразуемой матрицы зависит от ее размерности и величины коэффициента деформации a и имеет очень большие значения даже при малоразмерных преобразованиях, слегка отличающихся от ДПФ. Плохая обусловленность является неотъемлемым свойством матриц Вандермонда, связанных с реальными проблемами. Единственным исключением является случай отсутствия деформирования частотной оси, когда WDFT становится обычным ДПФ.

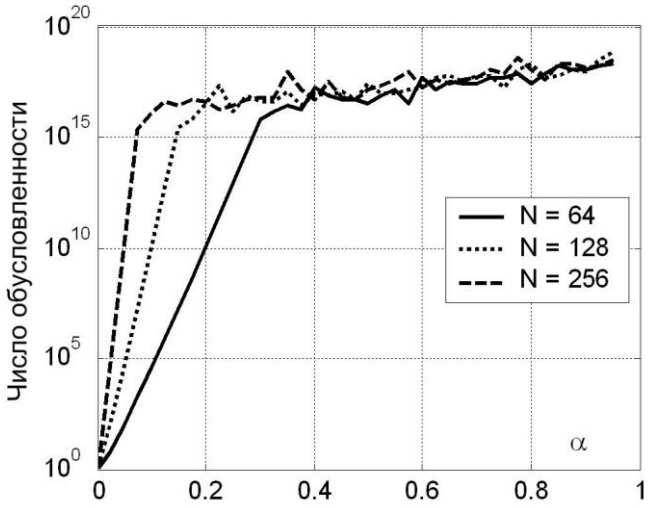


Рисунок 2.2 – Зависимость числа обусловленности матрицы от коэффициента деформации для различных размеров WDFT

Существует ряд методов для аппроксимации обращений плохо обусловленных матриц. Данные методы используют разложение по сингулярным числам матрицы (от англ. Singular Value Decomposition – SVD), которое определяется как

$$\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{H} = \sum_{i=1}^{N} \mathbf{u}_i \sigma_i \mathbf{v}_i^{H}, \tag{2.8}$$

где

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N], \\ \mathbf{V} &= [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N] \end{aligned} \tag{2.9}$$

матрицы с ортогональными столбцами $\mathbf{U}^{H}\mathbf{U}=\mathbf{V}\mathbf{V}^{H}=\mathbf{I}_N$ и

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_N) \tag{2.10}$$

является диагональной матрицей, состоящей из сингулярных чисел, отсортированных по убыванию $\sigma_1 \geq \cdots \geq \sigma_N \geq 0$. Столбцы матриц **U** и **V** – левый и правый – сингулярные векторы, H обозначает эрмитово транспонирование матрицы. Степень неполноты матрицы **D** (наличие некоторых линейных зависимостей между ее столбцами) проявляется в существовании почти нулевых сингулярных чисел σ_i. Распределения сингулярных значений для нескольких различных матриц WDFT проиллюстрированы на рисунке 2.3.

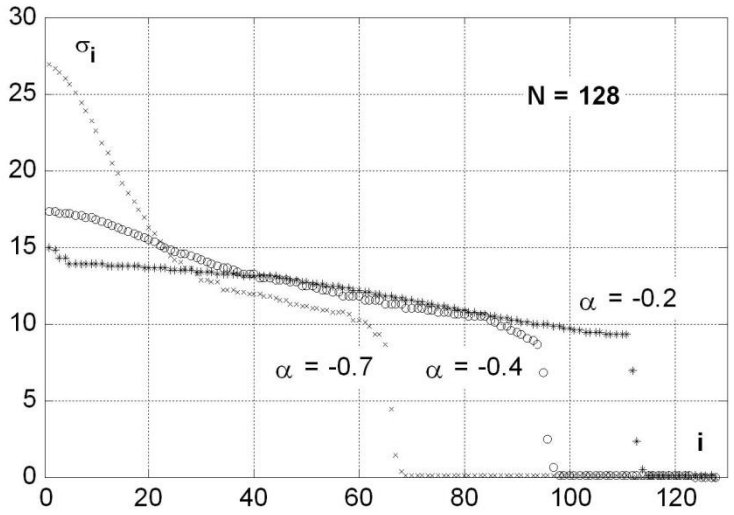


Рисунок 2.3 – Распределение сингулярных чисел как функция коэффициента деформации частотной шкалы для фиксированного размера преобразования $N = 128$

Как показывает рисунок 2.3, значения достаточно большого количества сингулярных чисел близки к нулю. Анализ распределения компонент SVD может дать много полезных пояснений плохой обусловленности матриц, а также они могут быть использованы для формирования псевдообратной матрицы:

$$\mathbf{D}^{\dagger} = \sum_{i=1}^{N} f_i \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^H. \qquad (2.11)$$

Через f_i обозначены так называемые коэффициенты ослабления фильтра и должны быть все равны единице для получения точного обращения матрицы. Теория регуляризации матриц рекомендует исключать влияние малых сингулярных чисел, ослабляя их вклад в (2.11). Это реализуется путем установки соответствующих значений f_i. Так как сингулярные векторы с наибольшими индексами имеют ярко выраженный колебательный характер, то это действие изменяет спектральный состав данных.

Основное различие между известными методами регуляризации заключается в выборе коэффициентов ослабления фильтра. В простейшем подходе усеченного SVD (от англ. Truncated SVD (TSVD)) сумма (2.11) ограничивается обычным отбрасыванием термов, имеющих сингулярные числа меньше определенного порога:

$$f_i = \begin{cases} 1 & \sigma_i > \lambda, \\ 0 & \text{в противном случае.} \end{cases} \qquad (2.12)$$

Менее радикальным является демпфированное SVD (от англ. Damped SVD (DSVD)), в котором коэффициенты ослабления фильтра постепенно изменяются в сторону нуля, обеспечивая более сглаженное отсечение:

$$f_i = \frac{\sigma_i}{\sigma_i + \lambda}. \qquad (2.13)$$

В обоих случаях параметр регуляризации λ должен быть выбран с осторожностью. Решение должно быть стабильным при спектральных изменениях, ограниченных определенным минимумом.

2.4 Оценка ошибки реконструкции сигнала

В зависимости от приложения, где применяется WDFT, ошибка реконструкции речевого сигнала (вычисление обратного WDFT) может влиять на качество синтезированного сигнала (число артефактов), например, в системах редактирования шумов окружающей среды, в речевом сигнале на основе психоакустически мотивированного правила взвешивания спектра зашумленной речи.

Используя матричную форму записи, вектор сигнала ошибки $\mathbf{d}$ можно определить как разность между оригинальным $\mathbf{x}$ и реконструированным сигналами $\hat{\mathbf{x}}$:

$$\mathbf{d} = \mathbf{x} - \hat{\mathbf{x}} = (\mathbf{I} - \mathbf{D}^{\dagger}\mathbf{D})\mathbf{x} \qquad (2.14)$$

Соответствующая мера в частотной области может быть выражена как спектральная плотность мощности (СПМ) сигнала ошибки:

$$S_{dd}(\omega) = \frac{1}{N} E\left\{\left|\mathbf{e}(\omega)^H \mathbf{d}\right|^2\right\} = \frac{1}{N}\mathbf{e}(\omega)^H \mathbf{R}_{\mathbf{dd}}\mathbf{e}(\omega)^H, \qquad (2.15)$$

где

$$\mathbf{e}(\omega) = [1 \quad e^{-j\omega 1} \quad e^{-j\omega 2} \quad \ldots \quad e^{-j\omega(N-1)}]^T \qquad (2.16)$$

является вектором-столбцом синусоидального базиса ДПФ, а $\mathbf{R}_{\mathbf{dd}}$ обозначает ковариационную матрицу сигнала ошибки.

Положим, что входной речевой сигнал $\mathbf{x}$ является случайным вектором с нулевым математическим ожиданием и известной ковариационной матрицей $\mathbf{R}_{\mathbf{xx}} = E\{\mathbf{x}\mathbf{x}^H\}$, тогда пусть $\mathbf{Q} = \mathbf{I} - \mathbf{D}^{\dagger}\mathbf{D}$ и матрица $\mathbf{R}_{\mathbf{dd}}$ может быть записана следующим образом:

$$\mathbf{R}_{\mathbf{dd}} = \mathbf{Q}\mathbf{R}_{\mathbf{xx}}\mathbf{Q}^H. \qquad (2.17)$$

Очевидно, что спектральное искажение (2.15) зависит от характеристик входного речевого сигнала и качества аппроксимации обратного WDFT. Теоретически точное обращение матрицы $\mathbf{D}$ возможно, но данное решение будет очень нестабильным вследствие большого числа обусловленности матрицы и

не найдет практического применения в системах обработки речи. Однако, если удается получить стабильное решение обратной задачи, мера величины ошибки (2.15) может быть использована для регулирования величины артефактов в синтезированном речевом сигнале. Так, если псевдообратная матрица $\mathbf{D}^{\dagger}$ определена с помощью техники SVD, представленной выше, то матрица $\mathbf{Q}$ может быть вычислена напрямую из выражения

$$\mathbf{Q} = \mathbf{I} - \mathbf{D}^{\dagger}\mathbf{D} = \mathbf{I} - \sum_{i=1}^{N} f_i \mathbf{v}_i \mathbf{v}_i^{H} . \qquad (2.18)$$

В случае перцептуальной деформации частотной шкалы выбор параметров регуляризации ограничивает влияние на уровень спектральных искажений. Другими словами, даже нестабильная аппроксимация обратной матрицы WDFT приводит к относительно высокой ошибке синтеза, а стабилизация решения только увеличит спектральные искажения. Например, положим, что коэффициент ослабления фильтра f_i вычисляется по методу DSVD (2.13) с параметром регуляризации λ = 0,001. СПМ ошибки WDFT-синтеза, вычисленная по формуле (2.15), показана на рисунке 2.4. В качестве входного сигнала в эксперименте использовался стационарный окрашенный гауссовский шум, который моделировался с априори заданной теплицевой ковариационной матрицей $(\mathbf{R}_{\mathbf{xx}})_{i,j} = p^{|i-j|}$ для p = 0,9. Можно заметить, что уровень искажений на заданной частоте зависит от расстояния между соседними WDFT-коэффициентами. Сигнал полностью восстанавливается только в точках преобразования (частотах, определенных выбором коэффициента α фазового звена (2.3)) и спектральные искажения особенно заметны в растянутых частотных диапазонах, в то время как в сжатых частотных диапазонах ошибка синтеза имеет приемлемый уровень.

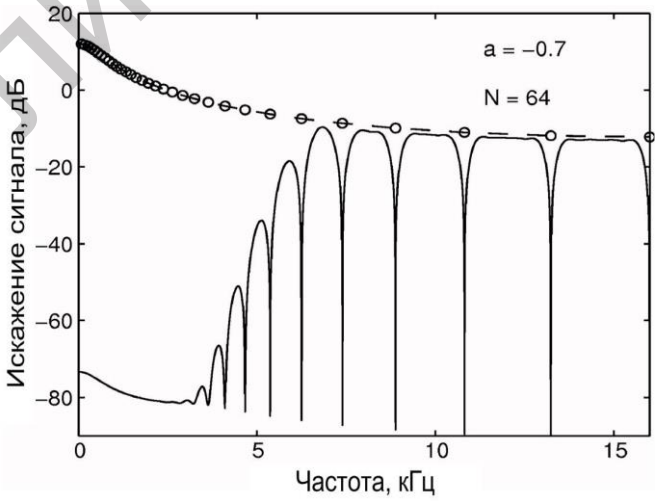


Рисунок 2.4 – СПМ оригинального сигнала (кружочки) и соответствующей ошибки синтеза (сплошная линия)

Этот эффект наблюдается также и на спектрограмме реконструированного речевого сигнала в виде узких спектральных «дыр», локализованных между точками преобразования (рисунок 2.5, б). Таким образом, частотная характеристика синтезированного речевого сигнала в некоторых деталях не восстанавливается. Заметим, что проявление данного эффекта в высокочастотной части частотного диапазона обусловлено перцептуальной деформацией частотной шкалы, т. е. коэффициент фазового звена $a < 0$, но ситуация меняется и эффекты ошибки WDFT-синтеза проявляются в низкочастотной части спектра, если $a > 0$. Однако данный случай не интересен для перцептуальных систем обработки речевых сигналов.

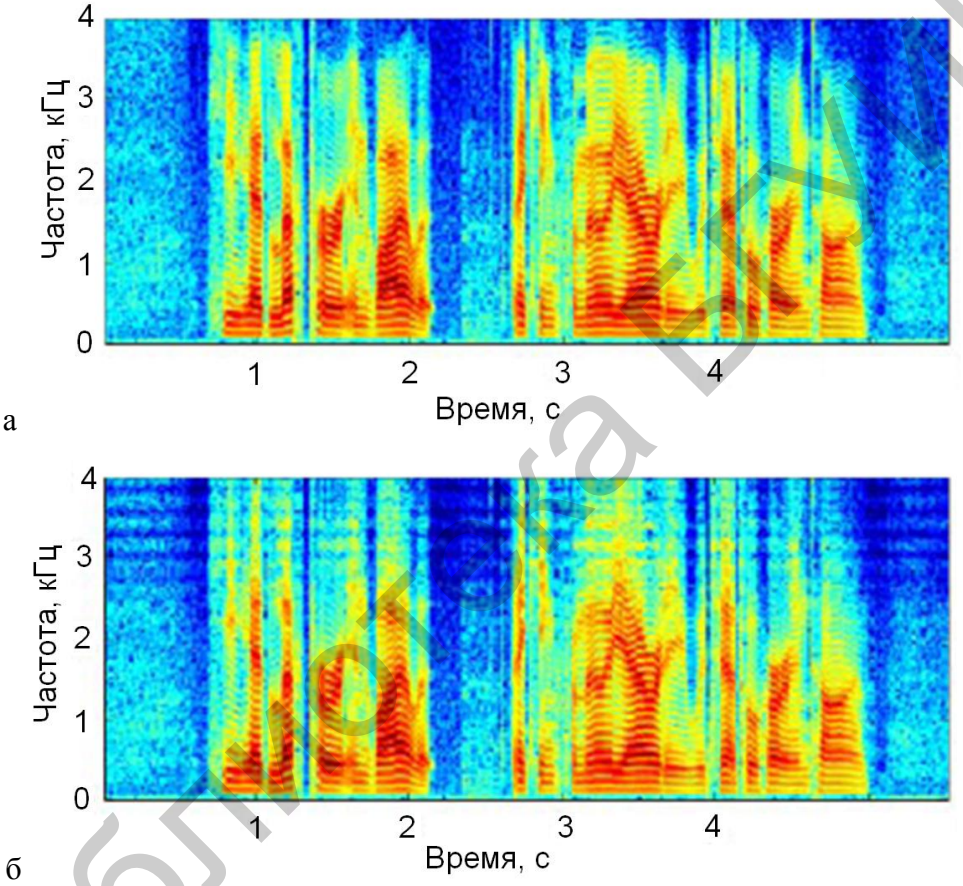


Рисунок 2.5 – Спектрограммы оригинального речевого сигнала (а) и реконструированного речевого сигнала (б)

Известно, что высокочастотный диапазон спектра речевого сигнала влияет на качество речи, в то время как низкочастотный – на разборчивость речи. Таким образом, при выборе параметров настройки WDFT необходимо искать некое компромиссное решение между требуемым частотным разрешением и ошибкой синтеза. Можно попытаться спроектировать фазовое звено с коэффициентом α, зависимым от времени, но данное решение имеет очень высокую вычислительную сложность и, более того, для широкополосного сигнала невозможно одновременно обеспечить его обработку с хорошей разрешающей способностью в низкочастотной части и перфективную реконструкцию его вы-

сокочастотных компонент, даже если удастся построить управляемое фазовое звено. Единственный путь минимизации спектрального искажения – это модификация матрицы преобразования **D** таким образом, чтобы количество почти нулевых сингулярных чисел было уменьшено.

2.5 Избыточный синусоидальный базис WDFT

Формирование избыточного неортогонального базиса может внести значительный вклад в коррекцию ошибки реконструкции, а именно, при подходящем выборе векторов синусоидального базиса можно модифицировать распределение сингулярных чисел для соответствующей матрицы преобразования. Новая WDFT-матрица **D** не является квадратной и количество строк $M > N$ увеличивается. Матричное представление WDFT с избыточным базисом может быть записано следующим образом:

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[M-1] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \hat{z}_0^{-1} & \cdots & \hat{z}_0^{-N+1} \\ 1 & \hat{z}_1^{-1} & \cdots & \hat{z}_1^{-N+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \hat{z}_{M-1}^{-1} & \cdots & \hat{z}_{M-1}^{-N+1} \end{bmatrix}}_{\mathbf{D}_{M\times N}} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}. \qquad (2.19)$$

Задача поиска обратной матрицы для данной прямоугольной матрицы $\mathbf{D}_{M\times N}$ может быть также решена использованием SVD-процедуры. Следует отметить, что стабильность новой матрицы выше, чем у матрицы стандартного WDFT. Избыточный базис с добавленными новыми векторами уменьшает эксцентричность SVD-эллипсоида, который является отображением единичной сферы в *N*-мерном пространстве. Соотношение между форматом WDFT с избыточным базисом и собственными числами для различных значений коэффициента деформации частотной шкалы α показано на рисунке 2.6. Видно, что если $M \to \infty$, то число обусловленности матрицы $\mathbf{D}_{M\times N}$ близко к единице.

С практической точки зрения необходимо выполнить два условия при формировании избыточного базиса WDFT. Во-первых, сохранить регулярность расположения коэффициентов преобразования в соответствии с деформацией частотной шкалы (например, шкалой барков). Во-вторых, размер нового синусоидального базиса должен быть как можно меньшим, чтобы вычислительная сложность WDFT оставалась приемлемой.

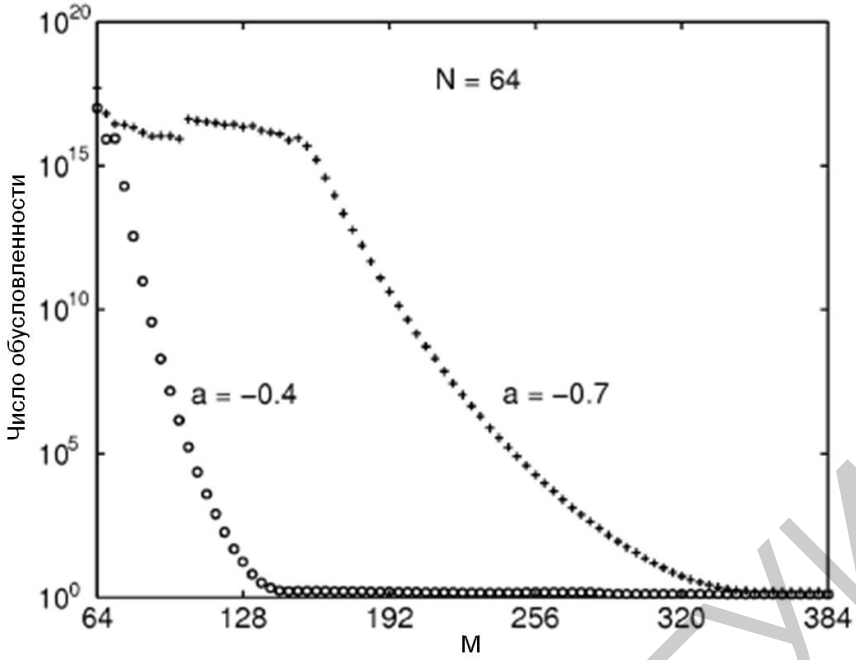


Рисунок 2.6 – Соотношение между форматом WDFT с избыточным базисом и собственными числами матриц $\mathbf{D}_{M\times N}$

Так как любое направление в пространстве комплексных векторов, определяемое базисным вектором, соответствует конкретному частотному диапазону, то WDFT можно представить как максимально децимированный банк фильтров. При этом *k*-я строка WDFT матрицы может рассматриваться как КИХ-фильтр с передаточной функцией, определяемой по следующему выражению:

$$H_k(z) = \sum_{n=-\infty}^{\infty} A(z_k) z^{-n} \quad k = 0, \ldots, N-1, \tag{2.20}$$

где $A(z)$ – фазовое звено первого порядка;

$H_k(z)$ – полосовой фильтр с центральной частотой

$$\hat{\omega}_k = 2\,\mathrm{arctg}\left(\frac{1+a}{1-a}\,\mathrm{tg}\left(\frac{\omega_k}{2}\right)\right), \quad \omega_k = \angle z_k . \tag{2.21}$$

и полосой $2\pi / N$.

АЧХ банка фильтров $H_k(z)$ для $k = 0, \ldots, N-1$ представлены на рисунке 2.7.

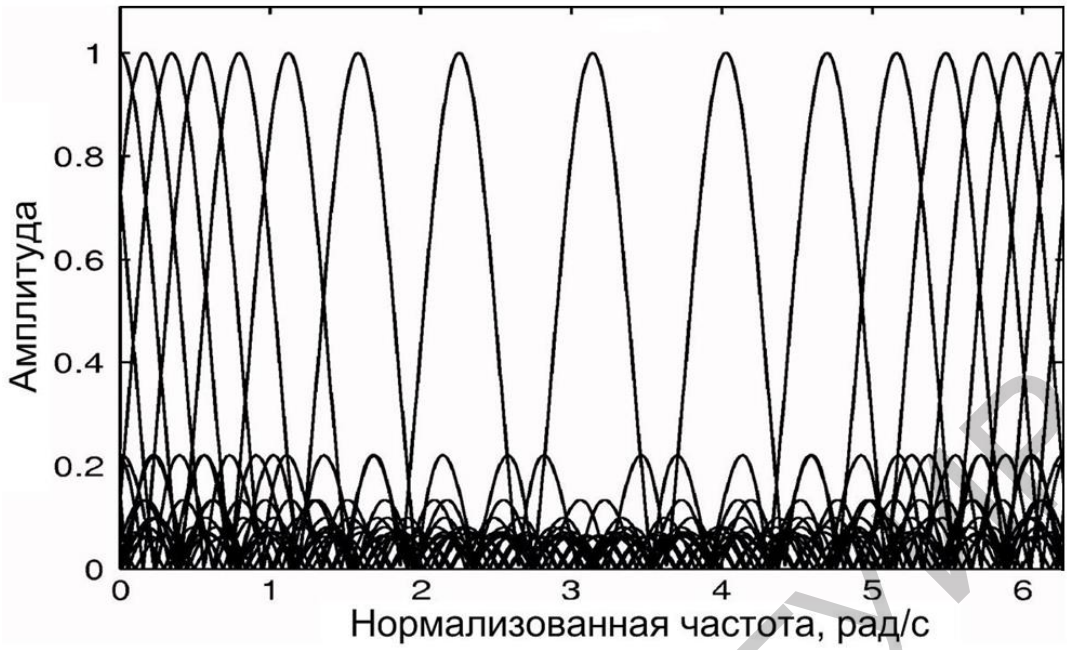


Рисунок 2.7 – АЧХ критически децимированного банка фильтров WDFT

Новый избыточный базис WDFT должен конструироваться из M векторов соответствующих импульсных характеристик КИХ-фильтров с центральными частотами

$$z_k = \exp(j\frac{2\pi k}{M}), k = 0,...,M-1, M > N, \tag{2.22}$$

регулярность которых на единичной окружности гарантирована, потому что вектор новых коэффициентов преобразования в точности совпадает со стандартным WDFT, но для формата преобразования M. Другими словами, размер нового избыточного базиса должен быть равен числу перекрывающихся полосовых фильтров (2.20), переопределенных для $k = 0,...,M-1$. Если коэффициент фазового звена в (2.3) $a < 0$, то максимальное расстояние между центральными частотами КИХ-фильтров составит

$$\Delta\omega_{\max} = \pi - 2\,\mathrm{arctg}\left(\frac{1+a}{1-a}\,\mathrm{tg}\left(\frac{\pi - 2\pi/M}{2}\right)\right). \tag{2.23}$$

Для получения того же частотного разрешения в высокочастотном диапазоне, что и у обычного ДПФ, расстояние между центральными частотами не должно быть больше чем $2\pi/N$. Подставляя $\Delta\omega_{\max} = 2\pi/N$ в (2.21) и решая относительно M, получаем, что

$$M = M_{\mathrm{opt}} = 2\pi\left[\pi - 2\,\mathrm{arctg}\left(\frac{1-a}{1+a}\,\mathrm{tg}\left(\frac{\pi - 2\pi/N}{2}\right)\right)\right]^{-1}. \tag{2.24}$$

Отметим, что для $\alpha = 0$ (нет деформации частотной шкалы) $M = N$, следовательно, $M \geq N$. Соотношение между размером избыточного базиса M и коэффициентом деформации α для заданного числа столбцов матрицы $\mathbf{D}_{M\times N}$ WDFT с избыточным базисом иллюстрируется на рисунке 2.8.

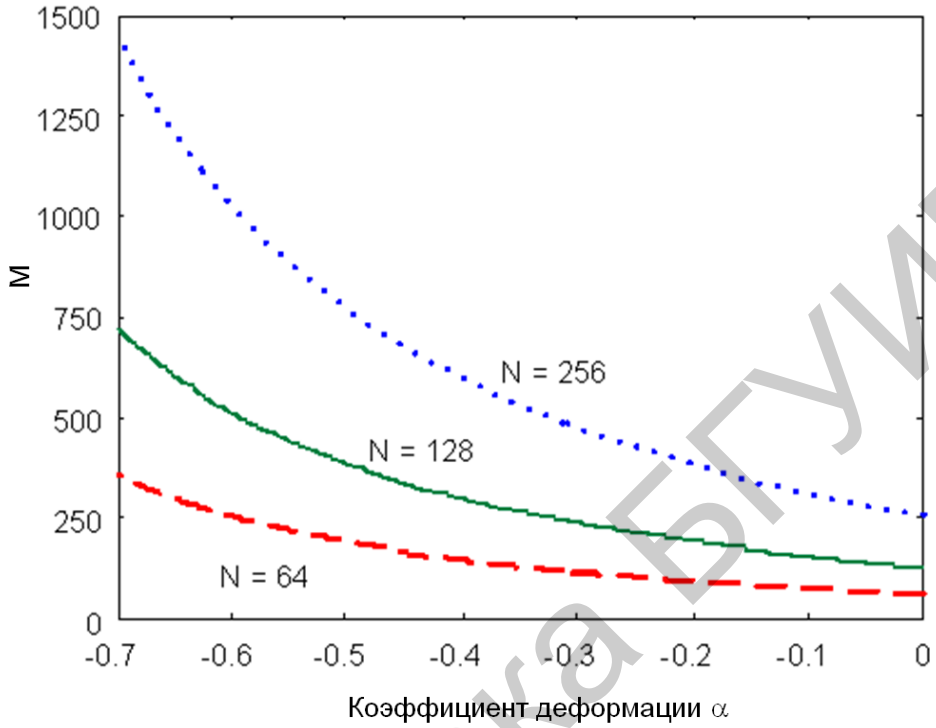


Рисунок 2.8 – Соотношение между размером избыточного базиса и коэффициентом деформации α для заданного числа столбцов матрицы $\mathbf{D}_{M\times N}$

Спектральные искажения (2.15) могут быть вычислены для прямоугольной матрицы так же, как и для квадратной, потому что продукт $\mathbf{D}^{\dagger}\mathbf{D}$ всегда $N \times N$-матрица. На рисунке 2.9 показаны СПМ входного сигнала и соответствующей ошибки синтеза, вычисленные для прямоугольной матрицы $\mathbf{D}_{M\times N}$ WDFT. Как видно, ошибка синтеза уменьшается при увеличении M и может не учитываться при $M \geq M_{\text{opt}}$ (меньше минус 75 дБ). На практике число строк матрицы $\mathbf{D}_{M\times N}$ можно слегка уменьшить, так как перекрытие частотных характеристик полосовых фильтров (2.20) выбиралось случайно и без всякого психоакустического критерия. Однако прослушивание речевых тестов показало, что спектральные искажения не слышны в реконструированном сигнале речи для $M \approx M_{\text{opt}}$.

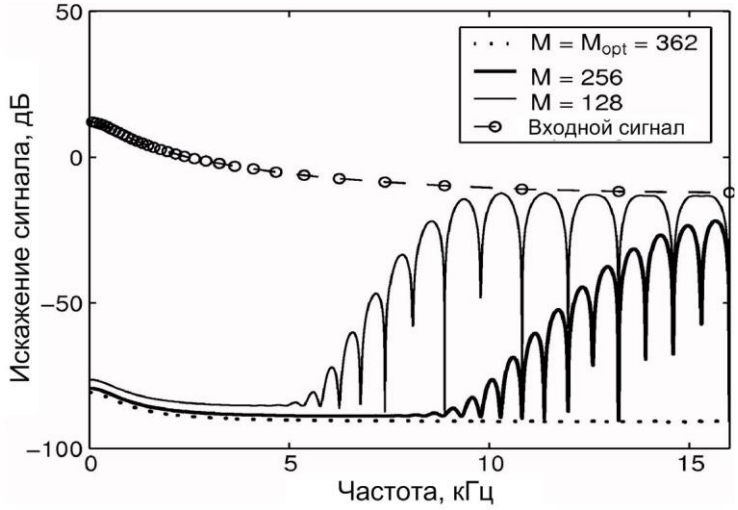


Рисунок 2.9 – СПМ входного сигнала (кружочки) и соответствующих спектральных искажений (сплошная линия) для разных форматов WDFT с избыточным базисом (N = 64)

2.6 Вычислительные аспекты WDFT

Алгоритм, по эффективности сравнимый с БПФ, не может быть построен для WDFT из-за асимметричности WDFT-матрицы. Тем не менее прямой алгоритм работы с комплексной матрицей может быть значительно оптимизирован. Наиболее усовершенствованный алгоритм, несмотря на сложность $O(N^2)$, использует факторизацию WDFT-матрицы в произведение трех матриц: действительной, ДПФ (вычисляемой через БПФ) и комплексной диагональной матрицы. Данный метод хорошо подходит для обработки изображений, где все данные поступают одновременно. В обработке речевых сигналов, где отсчеты следуют один за другим, может быть использована даже прямая реализация преобразования WDFT. Учитывая выражение (2.5), формула (2.6) может быть записана в следующем виде:

$$\begin{bmatrix} \hat{X}[0] \\ \hat{X}[1] \\ \vdots \\ \hat{X}[N-1] \end{bmatrix} = \sum_{n=0}^{N-1} \begin{bmatrix} A(z_0)^n \\ A(z_1)^n \\ \vdots \\ A(z_{N-1})^n \end{bmatrix} x[n]. \qquad (2.25)$$

Каждый элемент в данной сумме относится только к одному входному отсчету. Он может быть рассчитан при поступлении входного отсчета и результат аккумулируется. Коэффициенты преобразования успешно будут вычислены после N шагов. При этом вычислительная нагрузка, приведенная к входному отсчету, равна $O(N)$.

2.7 Аппроксимация шкалы барков

Для достижения высокого качества реконструированных сигналов в перцептуальных системах обработки речи требуется эффективная психоакустическая модель. В известной работе Джонстона психоакустическая модель основана на ДПФ: расчет ДПФ взвешенного временным окном сегмента сигнала, группировка коэффициентов преобразования в группы, соответствующие критическим частотным полосам, и расчет энергии в данных частотных полосах. Достижение приемлемого спектрального разрешения в критических частотных полосах, расположенных в низкочастотной части частотного диапазона, требует использования ДПФ с достаточно длинным временным окном. Поэтому концептуальная простота и эффективность нивелируются недостаточным временным разрешением, неприемлемым для анализа более тонкого феномена, такого как маскирование назад («pre-masking»).

Первый шаг при использовании *WDFT* в психоакустической модели – проектирование соответствующего фазового преобразования. Частотные коэффициенты z-преобразования должны быть распределены равномерно в перцептуальной шкале. Фазовое звено первого порядка достаточно хорошо аппроксимирует перцептуальную шкалу барков, при этом значение коэффициента фазового фильтра для заданной частоты дискретизации определяется по следующему выражению:

$$a_{\text{Bark}} = 0{,}1957 - 1{,}048 \cdot \left[\frac{2}{\pi} \operatorname{arctg}\left(0{,}07212 \frac{f_s}{1000} \right) \right]^{\frac{1}{2}}. \qquad (2.26)$$

Для частоты дискретизации $f_s = 16$ кГц коэффициент $a_{\text{Bark}} = -0{,}57827$.

Так как ширина критических частотных полос строго изменяется с их местоположением на частотной шкале, то различное количество коэффициентов преобразования ассоциируется с конкретной критической частотной полосой. В части А таблицы 2.1 количество коэффициентов в группах варьируется от 3 до 38 для ДПФ, в то время как для WDFT той же размерности не отдается предпочтения ни одной из полос, все коэффициенты преобразования распределены практически равномерно (часть Б).

WDFT в его оригинальной форме не сохраняет энергию сигнала в соответствующих частях единичной окружности до и после фазового звена. Вследствие этого каждый коэффициент WDFT должен масштабироваться в соответствии с коэффициентом $\sqrt{1-a^2}/(1-az)$, чтобы вычислить корректно уровни энергии в критических частотных полосах.

На рисунке 2.10 приведен пример обработки широкополосного сигнала, который состоит из голосового сообщения и музыкального фрагмента (рисунок 2.10, а). Распределение мощности в шкале барков для данного сигнала, вычисленное на основе спектрограммы WDFT ($a_{\text{Bark}} = -0{,}57827$) (рисунок 2.10, в), показывает, что оценка энергии в низкочастотном и высокочастотном диапазо-

нах может быть получена приблизительно одинаковой, чем если делать измерения по спектрограмме, полученной на основе ДПФ (рисунок 2.10, б).

Выбор WDFT малого формата с успехом может заменить ДПФ с большой длиной выборки благодаря тому, что коэффициенты WDFT-преобразования равномерно распределены в критических частотных полосах, поэтому в психоакустической модели на базе WDFT могут быть уравновешены как хорошее частотное, так и временное разрешение.

Таблица 2.1 – Сравнение распределения коэффициентов ДПФ и WDFT в критических частотных полосах

Крит. полоса	Часть А (размер ДПФ = 512, $f_s = 16$ кГц)			Часть Б (размер WDFT = 512, $f_s = 16$ кГц)		
	Диапазон коэффициентов	Кол-во	Диапазон частот, Гц	Диапазон коэффициентов	Кол-во	Диапазон частот, Гц
1	1–3	3	31–94	1–12	12	8–100
2	4–6	3	125–188	13–24	12	109–202
3	7–9	3	219–281	25–36	12	210–305
4	10–13	4	313–406	37–48	12	314–412
5	14–16	3	438–500	49–60	12	421–523
6	17–20	4	531–625	61–73	13	533–650
7	21–24	4	656–750	74–85	12	660–776
8	25–29	5	781–906	86–97	12	787–912
9	30–34	5	938–1063	98–110	13	923–1073
10	35–40	6	1094–1250	111–123	13	1086–1254
11	41–46	6	1281–1438	124–135	12	1269–1443
12	47–54	8	1469–1688	136–148	13	1460–1680
13	55–62	8	1719–1938	149–161	13	1700–1961
14	63–73	11	1969–2281	162–174	13	1985–2302
15	74–86	13	2313–2688	175–186	12	2331–2690
16	87–102	16	2719–3188	187–198	12	2726–3174
17	103–122	20	3219–3813	199–210	12	3220–3792
18	123–145	23	3844–4531	211–221	11	3851–4513
19	146–173	28	4563–5406	222–231	10	4588–5328
20	174–205	32	5438–6406	232–242	11	5419–6412
21	206–243	38	6438–7594	243–252	10	6520–7533
22	244–256	13	7625–8000	253–256	4	7650–8000

а

б

в

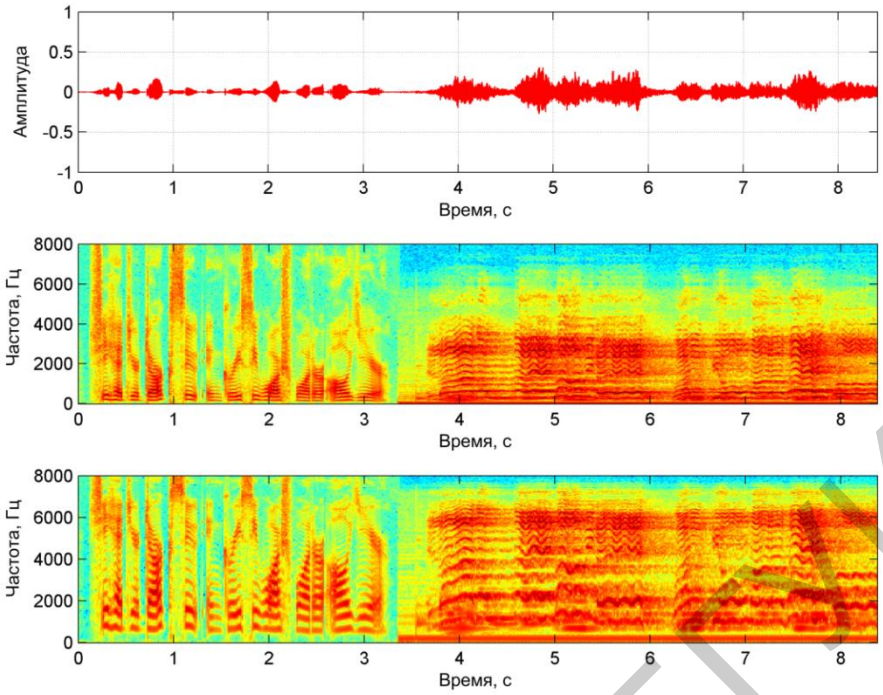


Рисунок 2.10 – Широкополосный сигнал во временной области (а) и его спектрограммы, полученные при помощи ДПФ (б) и WDFT (в)

3 ДИСКРЕТНОЕ ПРЕОБРАЗОВАНИЕ ФУРЬЕ С НЕРАВНОМЕРНЫМ ЧАСТОТНЫМ РАЗРЕШЕНИЕМ В ПЕРЦЕПТУАЛЬНЫХ СИСТЕМАХ ПОДАВЛЕНИЯ ШУМА В РЕЧИ

3.1 Общая схема системы шумоподавления

Большинство существующих систем подавления шума работают в частотной области, при этом используется хорошо известный подход спектрального вычитания или, другими словами, правило спектрального взвешивания (рисунок 3.1).

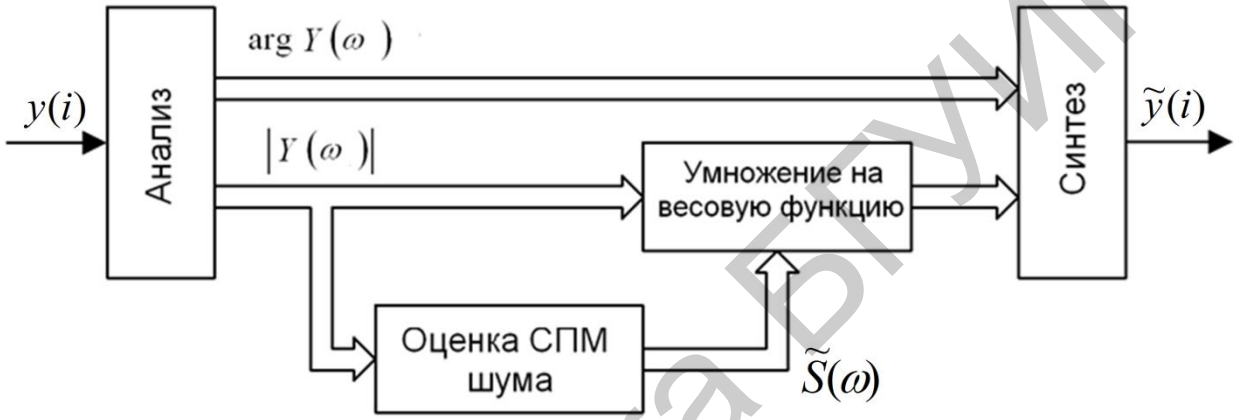


Рисунок. 3.1 – Схема подавления шума на основе спектрального взвешивания

Предполагается, что чистый речевой сигнал $s(i)$ и окружающий шум $n(i)$ статистически независимые и стационарные в широком смысле (i обозначает временной индекс), тогда оцифрованный зашумленный речевой сигнал может быть представлен следующим образом:

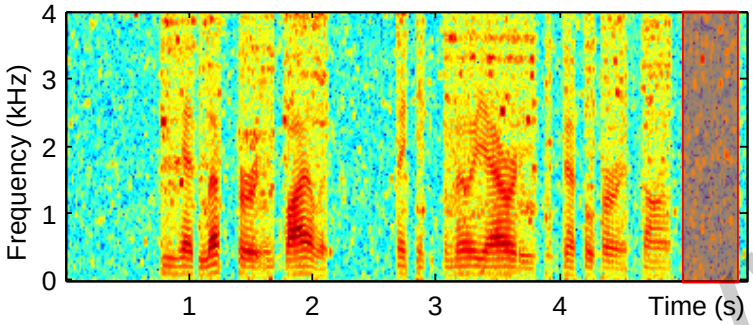
$$y(i) = s(i) + n(i). \qquad (3.1)$$

Входной сигнал $y(i)$ разбивается на перекрывающиеся фреймы из N последовательных отсчетов. Каждый фрейм умножается на временное окно и преобразуется в частотную область $Y(\omega)$, где $0 < \omega < 2\pi$. Уменьшение шума достигается умножением спектральных коэффициентов зашумленной речи на действительные коэффициенты весовой функции $H(\omega)$:

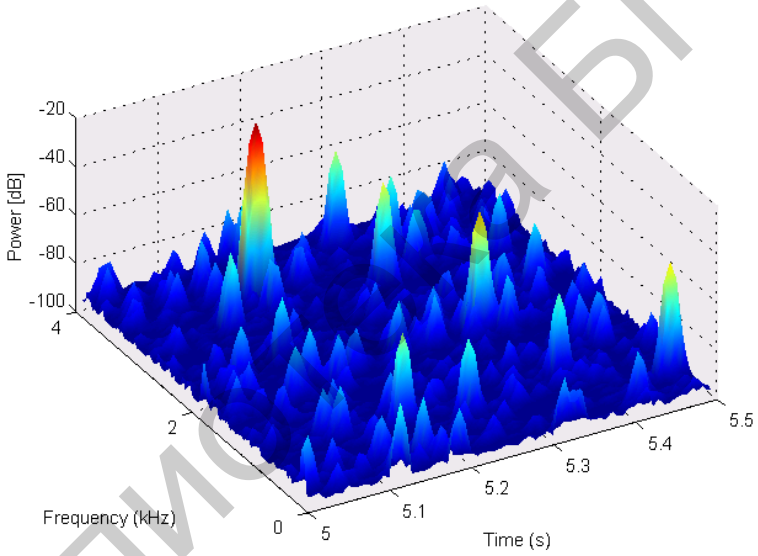
$$\tilde{S}(k) = H(\omega) \cdot Y(\omega), \quad 0 \le H(\omega) \le 1. \qquad (3.2)$$

Следует заметить, что фаза сигнала не изменяется. После модификации амплитуд спектр реконструированного системой сигнала $\tilde{S}(\omega)$ преобразуется обратно во временную область.

Несмотря на то что эти методы очень просты и легко реализуются, их слабым местом является остаточный шум, также известный как «музыкальный тон» (рисунок 3.2). Трехмерное изображение (рисунок 3.2, б) распределения мощности остаточного шума, соответствующее заштрихованной части спектрограммы (рисунок 3.2, а), очень наглядно демонстрирует недостаток данного подхода подавления шума в речевом сигнале.



а



б

Рисунок 3.2 – Спектрограмма выходного сигнала системы подавления шума (а), музыкальный тон (б)

3.2 Психоакустически мотивированное правило спектрального взвешивания

Возникает задача модификации взвешивающего правила на основе принципов психоакустики таким образом, чтобы оставить музыкальные тона немного ниже порога маскирования. Схема подавления шума на основе WDFT и психоакустически мотивированного правила спектрального взвешивания показана на рисунке 3.3.

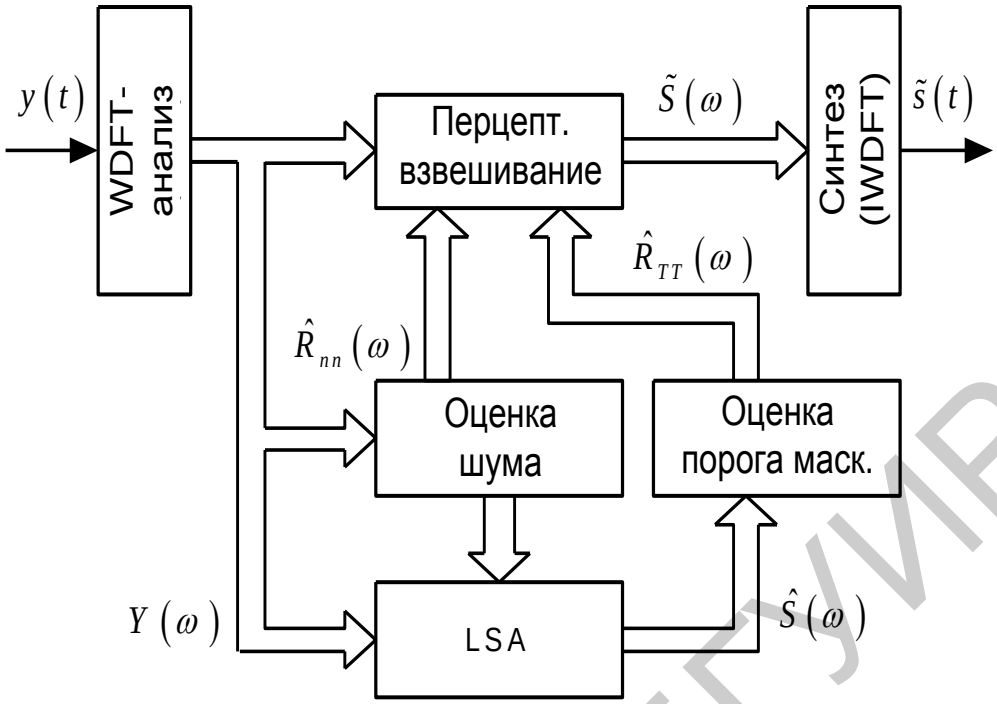


Рисунок 3.3 – Структура перцептуальной системы подавления шума на базе WDFT

Основной задачей данного подхода является не полное удаление шума, а ослабление слышимого шума до уровня комфортного шума. Это обусловлено тем, что в некоторых случаях, например во время диалога по телефону, полное удаление шума нежелательно. Для того чтобы сохранить характеристики окружающего шума, необходимо определить предпочтительный уровень остаточного шума ζ_n. Тогда разница между желаемым спектром реконструированного речевого сигнала и его оценкой может быть определена как

$$Q(\omega) = S(\omega) + \zeta_n N(\omega) - H(\omega)\left[S(\omega) + N(\omega)\right], \tag{3.3}$$

где $S(\omega)$ и $N(\omega)$ спектры чистой речи и шума соответственно;

$H(\omega)$ – оценка весовой функции.

Так как речь и шум предполагаются статистически независимыми, то спектральная плотность мощности (СПМ) разности (3.3) может быть выражена следующим образом:

$$R_{qq}(\omega) = \left[1 - H(\omega)\right]^2 R_{ss}(\omega) + \left[\zeta_n - H(\omega)\right]^2 R_{nn}(\omega), \tag{3.4}$$

или

$$R_{qq}(\omega) = R_{q_s q_s}(\omega) + R_{q_n q_n}(\omega), \tag{3.5}$$

Слагаемые выражения (3.5) соответствуют СПМ искажения речи и шума соответственно. Для минимизации суммы (3.5) в перцептуальном смысле необ-

ходимо, чтобы уровень искажений был невоспринимаемым на слух. В идеальном случае все искажения должны быть замаскированы. Тем не менее в большинстве реальных систем это требование не может быть удовлетворено, так как минимум $R_{qq}(\omega)$ может быть больше, чем СПМ порога маскирования $R_{TT}(\omega)$. Поэтому критерий минимизации формулируется как

$$R_{q_n q_n}(\omega) = R_{TT}(\omega). \tag{3.6}$$

Решая уравнение (3.6) относительно $H(\omega)$, определяем весовую функцию

$$H^{IND}(\omega) = \min\{1, \sqrt{\frac{R_{TT}(\omega)}{R_{nn}(\omega)}} + \zeta_n\}, \tag{3.7}$$

где *IND* означает неслышимое шумовое искажение (от англ. Inaudible Noise Distortion).

Легко заметить, что если мощность остаточного шума лежит ниже порога маскирования, тогда выражение под корнем больше единицы и речь не искажается, так как $H^{IND}(\omega) = 1$. В противном случае окружающий шум оптимально ослабляется до уровня, не воспринимаемого на слух человеком. Заметим, что оценки порога маскирования и СПМ шума необходимы только для вычисления взвешивающих коэффициентов. Простая модификация перцептуальной энтропии используется как базис для оценки порога маскирования чистой речи. Препроцессор оценки СПМ зашумленного речевого сигнала реализован на основе метода LSA (см. рисунок 3.3).

Применение WDFT в перцептуальной системе редактирования шума не только как базиса для определения модели маскирования, но и как инструмента спектрального анализа, позволяет добиться лучшего качества реконструированного речевого сигнала по сравнению с системами на основе ДПФ. Это объясняется тем, что весь процесс обработки осуществляется в перцептуальном домене с неравномерной частотной шкалой и нет необходимости в преобразованиях между разными частотными шкалами, что приводит к упрощению архитектуры системы. Более того, обработка речи, осуществляемая в критических частотных полосах, более точна в контексте психоакустического моделирования.

3.3 Отображение коэффициентов WDFT на критические частотные полосы

Проведенные исследования показали, что малоформатное WDFT может успешно заменить ДПФ с большой длиной выборки. Это оказалось возможным благодаря тому, что WDFT позволяет разместить частотные компоненты в соответствии с распределением критических частотных полос, поэтому в психо-

акустической модели на базе WDFT могут быть уравновешены как хорошее частотное, так и временное разрешение.

На основе свойств WDFT процедура оценки порога маскирования на базе общей психоакустической модели Джонстона была модифицирована. Первый шаг при использовании WDFT в психоакустической модели – проектирование соответствующего фазового звена. Частотные коэффициенты $_z$-преобразования должны быть представлены регулярно в перцептуальной области. Показано, что фазовое звено первого порядка достаточно хорошо аппроксимирует перцептуальную шкалу барков. При этом значение коэффициента фазового звена для частоты дискретизации f_s определяется по выражению (2.26). Для f_s = 8 кГц коэффициент $a_{\text{Bark}} = -0{,}4092$.

3.4 Оценка СПМ шума

Оценка дисперсии шума является ключевой задачей для многих систем повышения качества речевого сигнала. Наиболее общий подход основывается на статистических измерениях в течение речевых пауз с использование экспоненциального усреднения. Периоды речевых пауз определяются в зашумленном речевом сигнале на основе детекторов вокализованности речевого фрейма (от англ. Voice Activity Detector (VAD)). Следовательно, эффективность таких анализаторов шума строго зависит от уровня SNR и типа шума. В частности, они экстремально чувствительны к внезапным изменениям уровня шума. Другие, более устойчивые подходы основаны на методе статистического минимума, ключевая идея которого заключается в слежении за минимальным уровнем мощности шума для каждой спектральной компоненты на протяжении ряда фреймов сигнала. В случае уменьшения мощности шума быстрая корректировка оценки мощности шума здесь очевидна. Но в противоположной ситуации (мощность шума увеличивается) обновление результата оценки задерживается на целый период слежения. К сожалению, метод статистического минимума не совсем очевидный и получение оптимальной компенсационной процедуры достаточно затруднительно. В данной системе (см. рисунок 3.3) используется модифицированная процедура экспоненциального усреднения с контролем по минимуму энергии (от англ. Minima Controlled Recursive Averaging (MCRA)). Этот подход находится между обычным экспоненциальным усреднением и методом статистического минимума.

В оригинальном алгоритме MCRA оценка СПМ зашумленной речи основывается на методе периодограмм. В данном решении оценки СПМ сглаживаются в соответствующих критических частотных полосах – барках. А именно, предполагается, что для большинства практических применений при оценке СПМ зашумленной речи в критических частотных полосах справедливо утверждение

$$R_{yy}(k,l) \approx \hat{R}_{yy}(b,l), \qquad (3.8)$$

где $k \in K_b$;

l – номер обрабатываемого фрейма;

b – номер критической частотной полосы;

K_b – количество коэффициентов преобразования, попадающих в b-ю критическую частотную полосу;

k – номер коэффициента преобразования в b-й критической частотной полосе.

Во временной области усреднение осуществляется следующим образом:

$$\hat{R}_{yy}(b,l) = \alpha \hat{R}_{yy}(b,l-1) + (1-\alpha)\hat{S}_{yy}(b,l). \tag{3.9}$$

Оценка СПМ зашумленной речи на основе мгновенных оценок СПМ $|Y(k,l)|^2$ в b-й критической частотной полосе равна

$$\hat{S}_{yy}(b,l) = \frac{1}{\omega_{\mathrm{High},b} - \omega_{\mathrm{Low},b}} \sum_{b \in K_b} \Delta\omega_k |Y(k,l)|^2, \tag{3.10}$$

где $\omega_{\mathrm{Low},b}$, $\omega_{\mathrm{High},b}$ – границы b-й критической частотной полосы;

$\Delta\omega_k$ – полоса k-го фильтра в WDFT-спектре.

Заметим, что выражение (3.10) представляет аппроксимацию процесса интегрирования непрерывной функции СПМ по пространству неравномерно расположенных частотных отсчетов. Для стандартного ДПФ с постоянным частотным разрешением оценка (3.10) представляет собой среднее арифметическое мгновенных значений СПМ $|Y(k,l)|^2$. Сглаживание внутри критических частотных полос уменьшает дисперсию оценки СПМ. Хотя при этом разрешающая способность спектрального анализа также уменьшается. Однако это не скажется на окончательном результате редактирования шума в речевом сигнале, потому что большинство шумов имеют равномерные спектры и, более того, разрешающая способность уменьшается в соответствии с частотной избирательностью человеческого уха. Таким образом, подобно (3.8) можно утверждать, что при оценке СПМ шума в каждой критической частотной полосе выполняется равенство

$$R_{nn}(k,l) \approx \hat{R}_{nn}(b,l), \tag{3.11}$$

где $k \in K_b$.

Согласно MCRA-схеме, СПМ шума в b-й критической частотной полосе и для l-го фрейма сигнала оценивается следующим образом:

$$\hat{R}_{nn}(b,l) = \tilde{\alpha}_n(b,l)\hat{R}_{nn}(b,l-1) + \left(1-\tilde{\alpha}_n(b,l)\right)\hat{S}_{yy}(b,l), \tag{3.12}$$

где $\tilde{\alpha}_n(b,l)$ – зависящий от времени коэффициент сглаживания, являющийся функцией оценки вероятности $p(b,l)$ вокализованности (присутствия речевой активности) *l*-го фрейма речевого сигнала:

$$\tilde{\alpha}_n(k,l) = \alpha_n + (1-\alpha_n)\, p(b,l), \qquad 0 < \alpha_n < 1, \tag{3.13}$$

Коэффициент сглаживания α_n контролирует общую способность слежения при оценке СПМ шума. Вероятность $p(b,l)$ вокализованности *l*-го фрейма рассчитывается на основе экспоненциального усреднения решений $I(b,l)$ несложного детектора речевой активности (VAD), реализованного на подходе минимума статистики:

$$p(b,l) = \alpha_p\, p(b,l-1) + (1-\alpha_p)\, I(b,l), \quad 0 < \alpha_p < 1, \tag{3.14}$$

где α_p – коэффициент сглаживания.

Решение VAD $I(b,l)$ интерпретируется как бинарный результат сравнения СПМ $\hat{R}_{yy}(b,l)$ зашумленного речевого сигнала и оценки СПМ детектора минимума статистики $R_{\min}(b,l)$ в *b*-й критической частотной полосе *l*-го фрейма сигнала:

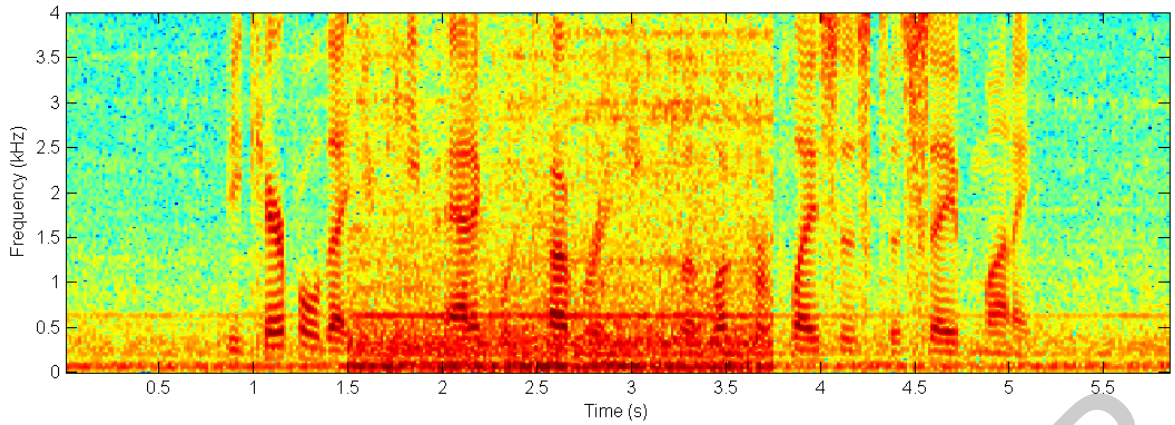
$$I(b,l) = \begin{cases} 1, & \text{где } \hat{R}_{yy}(b,l) \geq B \cdot R_{\min}(b,l), \\ 0 & \text{иначе.} \end{cases} \tag{3.15}$$

Здесь параметр *B* выбирается эмпирически как положительная константа (обычно 2< *B* <3). Оценка детектора минимума статистики получается согласно правилу:

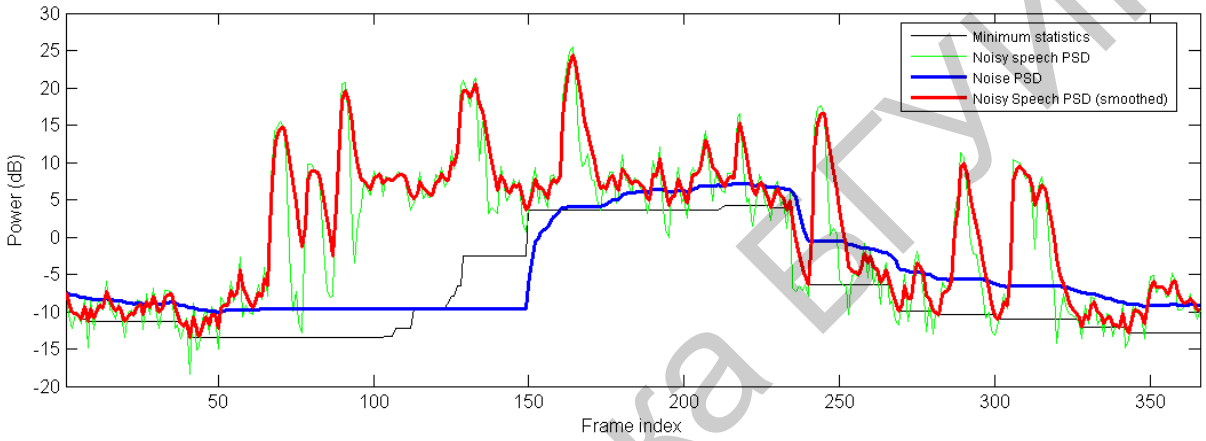
$$R_{\min}(b,l) = \min\left\{\hat{R}_{yy}(b,l-T), \hat{R}_{yy}(b,l-T+1), \ldots, \hat{R}_{yy}(b,l)\right\}, \tag{3.16}$$

где *T* обозначает длину окна поиска минимума.

На рисунке 3.4 показан пример слежения в зашумленном речевом сигнале за изменением СПМ, у которого внезапно изменился уровень мощности шума. В данном эксперименте параметр *T* = 60 фреймам. Как можно видеть из рисунка 3.4, б, отклик алгоритма MCRA на скачок мощности шума имеет заметное опоздание за счет длины окна поиска минимума. В случае уменьшения уровня мощности, как показывает эксперимент, алгоритм реагирует незамедлительно. На рисунке 3.4, а представлена спектрограмма данного зашумленного речевого сигнала.



а



б

Рисунок 3.4 – Пример слежения за изменением СПМ на основе алгоритма MCRA

3.5 Оценка порогов маскирования

Основные шаги оценки порога маскирования следующие: расчет энергии критических частотных полос по WDFT-спектру мощности; свертка с функцией распространения; вычитание относительного смещения порога; нормализация и сравнение с абсолютным порогом слышимости.

Дискретное представление энергии критической частотной полосы может быть выражено как взвешенная сумма компонент спектра мощности:

$$E_b = \frac{1}{2\pi} \sum_{k \in K_b} \Delta\omega_k \left| \hat{S}(\omega_k) \right|^2, \qquad (3.17)$$

где k – номер коэффициента преобразования (спектрального отсчета);

$\hat{S}(\omega_k)$ – оценка спектра оригинального речевого сигнала.

На следующем шаге вычисляется свертка энергий критических частотных полос E_b с функцией распространения по базилярной мембране SF_b для учета распространения маскирования:

$$C_b = E_b \odot SF_b, \tag{3.18}$$

где $\odot$ – знак свертки.

Относительное смещение порога в каждой критической частотной полосе рассчитывается с использованием меры тональности. Для определения типа маскера («шумоподобный» или «тональный») используется мера пологости спектра, которая определяется как отношение среднего геометрического μ_{gb} к среднему арифметическому μ_{ab} спектральных компонент внутри критической частотной полосы b:

$$SFM_b = 10\log_{10}\left(\frac{\mu_{gb}}{\mu_{ab}}\right), \text{ дБ.} \tag{3.19}$$

Получение данной оценки может быть затруднено в случае обычных моделей с малым разрешением, так как количества ДПФ-коэффициентов в низкочастотных полосах обычно недостаточно для эффективного расчета *SFM*. В подобных случаях *SFM* рассчитывается для всего спектра или предопределенных тональностей для каждой критической частотной полосы. В случае WDFT коэффициенты преобразования группируются равномерно в критических частотных полосах, следовательно, *SFM* может быть определена для каждой полосы отдельно.

Смещение порога маскирования O_b оценивается по следующему выражению:

$$O_b = \text{ton}_b \cdot (14{,}5 + b) + (1 - \text{ton}_b) \cdot 5{,}5, \text{ дБ,} \tag{3.20}$$

где ton_b – индекс тональности, определяемый как

$$\text{ton}_b = \min\left(\frac{SFM_b}{-60 \text{ дБ}}, 1\right). \tag{3.21}$$

Для получения энергии порога маскирования относительное смещение вычитается из свернутого спектра критической полосы:

$$P_{TT,b} = 10^{(\log_{10} C_b - O_b/10)}, \tag{3.22}$$

а затем осуществляется нормализация и сравнение с абсолютным порогом слышимости.

Таким образом, оценки СПМ шума так же, как и оценки порога маскирования, вычисленные на их основе, на обрабатываемом фрейме речевого сигнала постоянны в критических частотных полосах, а следовательно, и соответствующие спектральные взвешивающие коэффициенты внутри критических ча-

стотных полос также постоянны. Это свойство не только упрощает схему обработки, но также ослабляет генерацию «музыкальных тонов».

3.6 Оценка качества системы подавления шума

Эксперименты по оценке качества реконструированного системой речевого сигнала проводились для следующих параметров настройки алгоритмов: частота дискретизации входного сигнала – 8 кГц, размер фрейма $N = 256$ отсчетов (32 мс), обработка осуществляется с 50%-м перекрытием фреймов, временное окно Хеннинга. В порядке сокращения вычислительной сложности размер окна анализа WDFT с избыточным базисом был сокращен до 128 отсчетов. Оптимальный формат WDFT для $N = 128$ и коэффициента фазового звена $a_{Bark} \approx -0{,}4$ находится около 300. Данную величину можно немного уменьшить благодаря перцептуальному маскированию искажений в речевом сигнале. Таким образом, размер матрицы $\mathbf{D}_{M\times N}$ WDFT с избыточным базисом был выбран 256×128.

Тестирование системы подавления шума осуществлялось на 8 озвученных предложениях русского языка. Длина речевых сегментов колебалась от 5 до 8 с. К каждому сегменту чистого речевого сигнала добавлялся «цветной» шум. Сегментное соотношение сигнал/шум (SEGSNR) варьировалось от минус 5 до 20 дБ. Использовались следующие объективные показатели качества: «Искажение речи» (в децибелах), «Кепстральное расстояние», «Ослабление шума» в децибелах. Оценка «Искажение речи» (в децибелах) это SEGSNR, где под шумом понимается разница между оригинальным (чистым) речевым сигналом и выходным сигналом системы подавления шума. Высокое значение данного SEGSNR указывает на малые искажения речи. Большое значение показателя «Кепстральное расстояние» говорит о наличии в выходном сигнале системы сильных артефактов речи. Под оценкой «Ослабление шума» понимается отношение мощности входного зашумленного речевого сигнала к мощности выходного сигнала системы, в котором шум был ослаблен. Однако объективные показатели качества слабо коррелируют с результатами оценок на основе восприятия человеком речи. Поэтому при тестировании использовался показатель «Перцептуальные искажения» на основе модифицированного метода искажений спектра барков (от англ. Modified Bark Spectral Distortion (MBSD)), который определяется как разница между восприятием чистой речи и синтезированной системой.

Экспериментальные результаты проиллюстрированы на рисунке 3.5 для системы подавления шума на основе стандартного ДПФ, WDFT «чистого» и с избыточным базисом. Результаты для WDFT с избыточным базисом значительно лучше для оценки «Ослабление шума», чем для ДПФ и «чистого» WDFT. Оценки показателей «Искажение речи» и «Кепстральное расстояние» системы на базе ДПФ и WDFT с избыточным базисом приблизительно одинаковые и намного лучше, чем для «чистого» WDFT. Это обусловлено значительным уменьшением ошибки синтеза у алгоритма WDFT с избыточным базисом по

сравнению с «чистым» WDFT. Анализ показателя MBSD показывает, что качество системы на основе WDFT с избыточным базисом не намного выше по сравнению с системами подавления шума на базе ДПФ и «чистого» WDFT. Это подтверждает предположение, что генерируемые искажения системой синтеза на основе «чистого» WDFT только отчасти слышны.

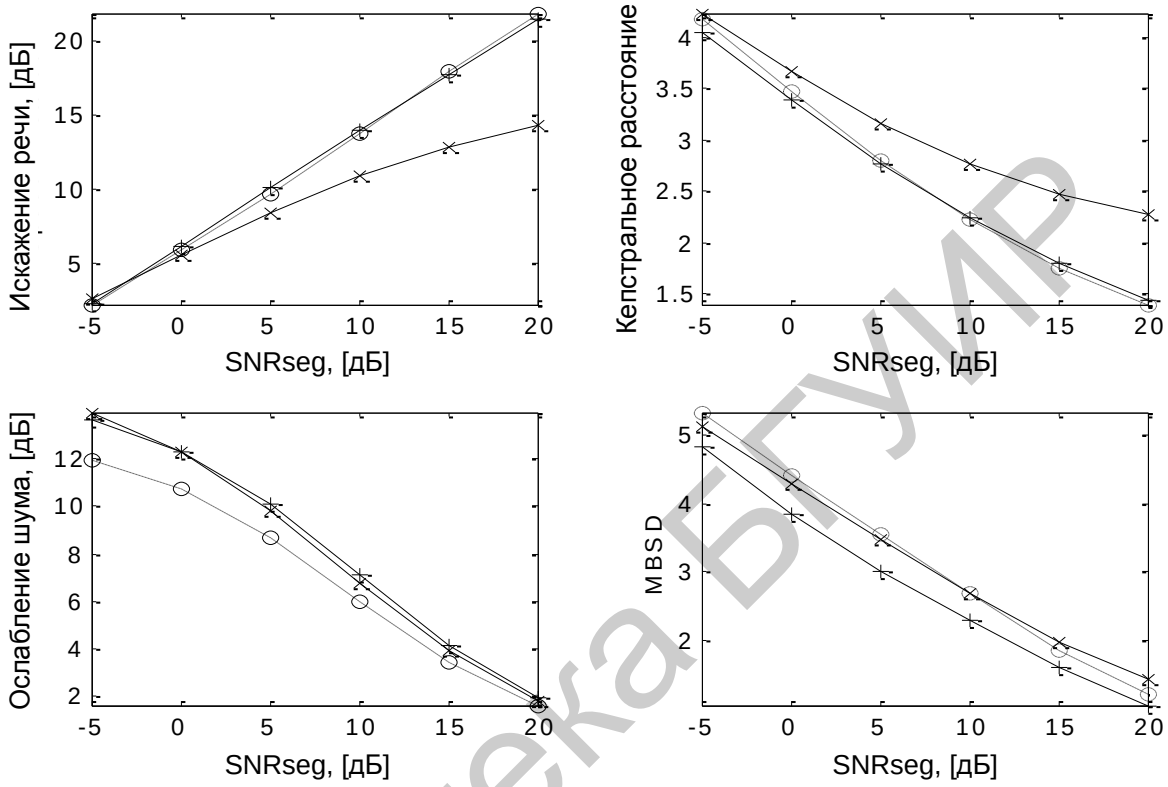
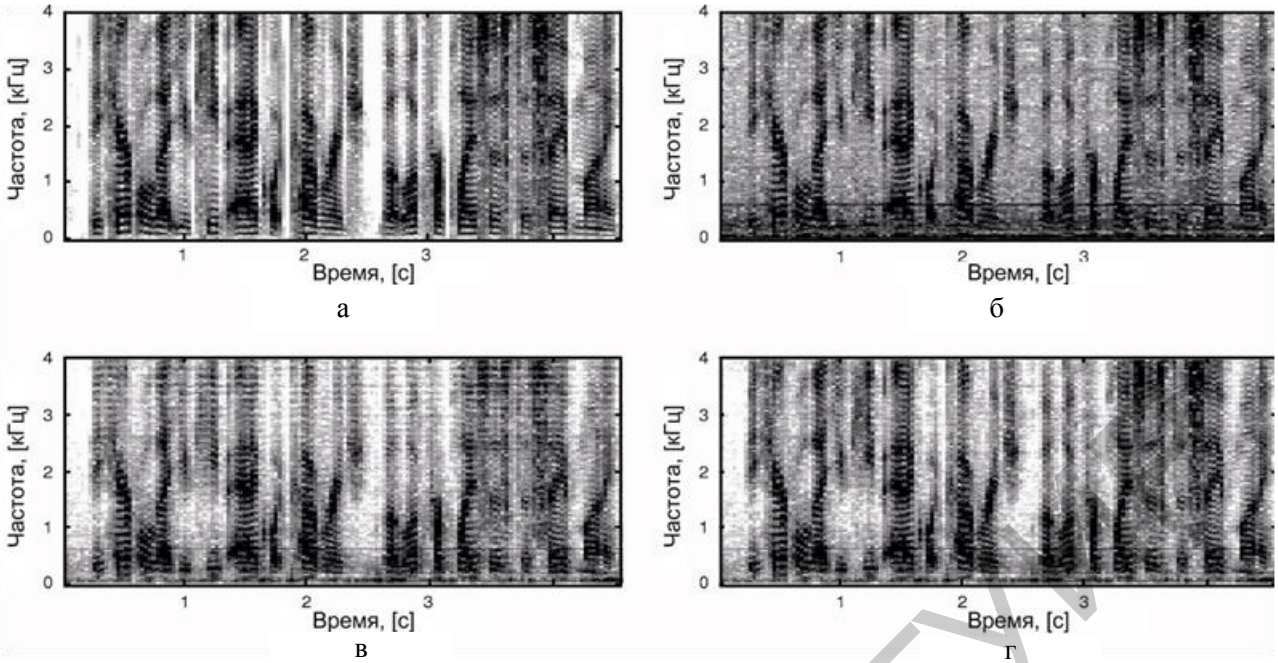


Рисунок 3.5 – Оценка качества систем на базе ДПФ (o), WDFT (×), WDFT с избыточным базисом (+)

Для анализа распределения мощности остаточного шума и искажений речи по частотному диапазону использовались спектрограммы (рисунок 3.6). Как видно, нет заметной разницы между системами в низкочастотном диапазоне. Во всех случаях шум окружающей среды сильно ослабляется и музыкальные тона не воспринимаются благодаря относительно высокому предопределенному уровню остаточного шума (3.3). Однако для системы на основе WDFT с избыточным базисом, в отличие от решения с «чистым» WDFT, ясно видно, что высокочастотные искажения не генерируются и высшие гармоники в реконструированном системой речевом сигнале близки к оригинальному сигналу.

Дальнейшие работы в данном направлении нацелены на уменьшение вычислительной сложности и разработку WDFT с высоким частотным разрешением на основе психоакустической модели (возможно для ERB-модели).



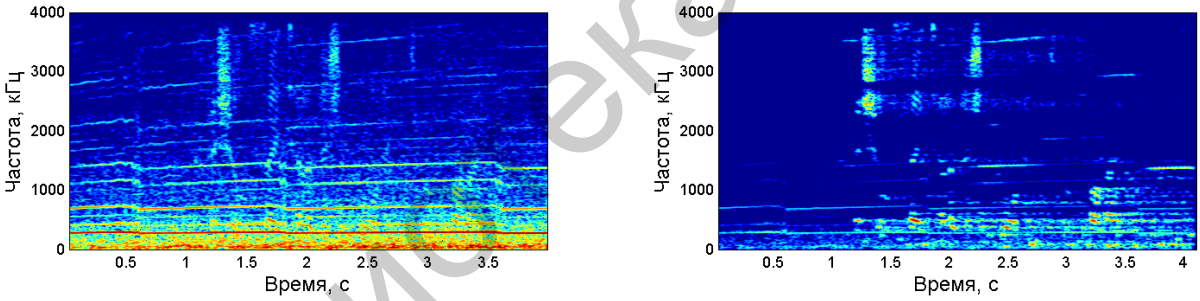
а – оригинальный речевой сигнал;
б – зашумленный речевой сигнал (шум мотора с SEGSNR = 5 дБ);
в – выходной сигнал системы на основе «чистого»WDFT;
г – выходной сигнал системы на базе WDFT с избыточным базисом

Рисунок 3.6 – Спектрограммы

4 АЛГОРИТМ ОЧИСТКИ РЕЧЕВОГО СИГНАЛА ОТ СЛОЖНЫХ ПОМЕХ ПУТЕМ ФИЛЬТРАЦИИ В МОДУЛЯЦИОННОЙ ОБЛАСТИ

4.1 Общие теоретические сведения

Одним из эффективных подходов к задаче шумоочистки является использование методов параметрического моделирования для подавления специальных шумов (помех), представляющих собой смесь нестационарных периодических и стохастических составляющих, зависящих от скорости источника шума. Такого рода шумы создаются различными вращательными механизмами, такими как турбины и двигатели внутреннего сгорания. Описываемый способ основывается на спектральном вычитании, однако позволяет учитывать нестационарную природу шума. Общепринятый метод спектрального вычитания для таких шумов имеет лишь ограниченную применимость, поскольку оценка СПМ связана со скоростью вращения и, таким образом, непрерывно изменяется. На рисунке 4.1 показан пример подавления нестационарного шума при помощи традиционного спектрального взвешивания. Из-за того что параметры шума изменяются быстро, выполнение качественного шумоподавления не обеспечивается. Видно, что большая часть исходной помехи осталась в сигнале, и, кроме того, исходный сигнал заметно деградировал.



а – исходный сигнал; б – обработанный сигнал

Рисунок 4.1 – Подавление нестационарного шума методом спектрального взвешивания

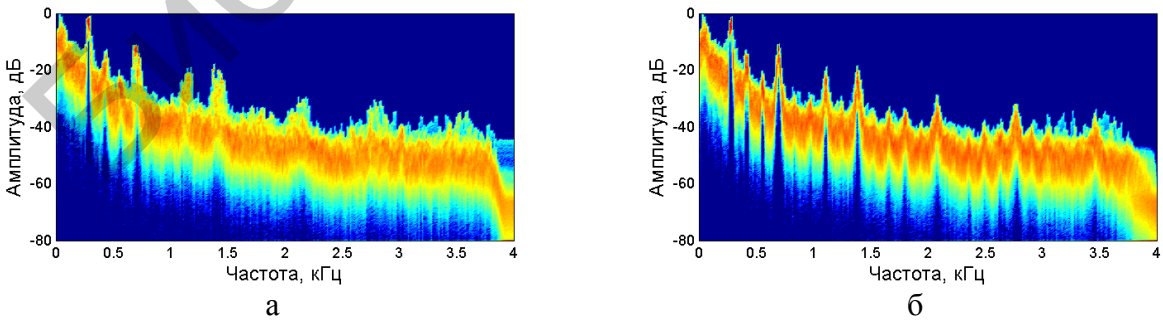
Проблема обработки таких сигналов характеризуется следующими факторами: неустойчивость параметров шума, что требует разработки алгоритмов слежения; очень низкое соотношение сигнал/шум, которое может быть менее минус 10 дБ; гибридная структура шума, которая представляет собой смесь детерминированных (тональных или узкополосных) и случайных (непериодических широкополосных) составляющих.

Среди применяемых подходов для решения данной задачи используются спектральное вычитание, фильтрация Винера и апостериорный анализ MAP (Maximum a Posteriori Method). Большинство методов используют дополнительную внешнюю информацию, такую как оборотная скорость двигателя, для

того чтобы определять текущую частоту гармоник шумового сигнала. При изменении скорости вращения и узкополосные, и широкополосные составляющие шума смещаются в частотной области, однако фильтрация, согласованная со скоростью вращения, выполняется при помощи режекторных фильтров и применяется только к узкополосным составляющим. Режекторная фильтрация позволяет подавить отдельные гармоники и широко используется в таких задачах, поскольку применение стандартных методов обработки сигнала в частотной области, таких как преобразование Фурье либо банки фильтров, затруднено в связи с нестационарностью наблюдаемого процесса. Широкополосные шумовые составляющие обрабатываются вне зависимости от скорости вращения, что является серьезным ограничением используемых методов, поскольку приводит к появлению избыточного шумового остатка в обработанном сигнале.

Во многих случаях возможно компенсировать изменение СПМ при помощи адаптивной дискретизации сигнала, согласованной со скоростью вращения. Сигнал представляется в измененном масштабе времени, что делает параметры шума более устойчивыми и доступными для оценки. Уменьшение вариативности СПМ шума позволяет выполнить более точную оценку статистики шума и усиливает эффект шумоподавления. Также становится возможной обработка сигнала при помощи сравнительно длинных окон анализа и использование дискретного преобразования Фурье для спектрального взвешивания. На рисунке 4.2 показано распределение значений амплитудного спектра шума двигателя автомобиля без адаптивной дискретизации и с адаптивной дискретизацией. Видно, что спектральные компоненты шумового сигнала становятся более устойчивыми после адаптивной дискретизации, согласованной со скоростью вращения.

Для оценки текущей скорости вращения можно использовать внешние измерители скорости, и выполнять анализ зашумленного сигнала – для оценки частоты основного тона шумового сигнала. Учитывая, что шум, как правило, состоит из смеси квазипериодических и случайных компонент, целесообразно выполнять спектральное взвешивание в два этапа, ослабляя узкополосные и широкополосные компоненты отдельно.



а б

а – исходный шумовой сигнал; б – шумовой сигнал с адаптивной дискретизацией

Рисунок 4.2 – Распределение значений амплитудного спектра шума

4.2 Метод спектрального взвешивания

Спектральное взвешивание представляет собой метод восстановления СПМ либо амплитудного спектра из сигнала, зарегистрированного с аддитивным шумом, путем вычитания средней оценки спектра шума. Спектр шума обычно оценивается и обновляется в периоды, когда полезный сигнал отсутствует, а присутствует только шум. Метод подразумевает, что шум является стационарным, либо медленно изменяющимся процессом, и не изменяется существенно между периодами обновления параметров.

Модель зашумленного сигнала представляется в следующем виде:

$$y(m) = x(m) + n(m), \tag{4.1}$$

где $y(m)$ – зашумленный сигнал;
$x(m)$ – чистый сигнал;
$n(m)$ – шум;
m – номер отсчета (дискретное время).

В частотной области модель сигнала описывается следующим образом:

$$Y(\omega) = X(\omega) + N(\omega), \tag{4.2}$$

где $Y(\omega)$ – преобразование Фурье зашумленного сигнала;
$X(\omega)$ – преобразование Фурье исходного чистого сигнала,
$N(\omega)$ – преобразование Фурье шума;
ω – частота.

При спектральном вычитании входной сигнал разделяется на сегменты фиксированной длины, взвешивается при помощи оконной функции и переводится в частотную область при помощи дискретного преобразования Фурье. Операция взвешивания на оконную функцию описывается в частотной области следующим образом:

$$Y_w(\omega) = W(\omega) \cdot Y(\omega) = X_w(\omega) + N_w(\omega). \tag{4.3}$$

Для краткости далее предполагается, что используется сигнал, взвешенный на оконную функцию, и индекс w опускается. Спектральное взвешивание описывается следующим выражением:

$$\left|\hat{X}(\omega)\right|^b = |Y(\omega)|^b - \alpha\overline{|N(\omega)|^b}, \tag{4.4}$$

где $\left|\hat{X}(\omega)\right|^b$ – полученная оценка спектра исходного чистого сигнала $|X(\omega)|^b$;
$\overline{|N(\omega)|^b}$ – средняя по времени оценка спектра шума.

Предполагается, что шум является в широком смысле стационарным случайным процессом. Для вычитания амплитудного спектра используется $b = 1$, а для вычитания спектра мощности $b = 2$. Параметр α определяет величину части шума, которая вычитается из зашумленного сигнала. Для полного вычитания $\alpha = 1$, а для вычитания с избытком $\alpha > 1$. Средняя оценка спектра шума вычисляется на фрагментах, где отсутствует полезный сигнал, следующим образом:

$$\overline{|N(\omega)|^b} = \frac{1}{K}\sum_{i=0}^{K-1}|N_i(\omega)|^b, \tag{4.5}$$

где $|N_i(\omega)|$ – спектр фрейма шума с индексом i;

K – число фреймов, где присутствует только шум.

На практике удобно оценивать средний спектр шума при помощи экспоненциального усреднения:

$$\overline{|N_i(\omega)|^b} = \rho\overline{|N_i(\omega)|^b} + (1-\rho)\overline{|N_i(\omega)|^b}, \tag{4.6}$$

где коэффициент усреднения ρ принимает значения из диапазона от 0,85 до 0,99.

Для восстановления обработанного сигнала во временной области оценка амплитудного спектра $\left|\hat{X}(\omega)\right|$ объединяется с фазовым спектром зашумленного сигнала и затем вычисляется дискретно-обратное преобразование Фурье:

$$\hat{x}(m) = \sum_{k=0}^{N-1}\left|\hat{X}(k)\right| e^{j\theta_Y(k)} e^{-j\frac{2\pi}{N}km}, \tag{4.7}$$

где $\theta_Y(k)$ – фаза частотной составляющей $Y(k)$ зашумленного сигнала;

N – длина фрейма обработки.

Процедура восстановления сигнала основывается на предположении, что аддитивный шум искажает, главным образом, амплитудный спектр и что искажения, вносимые в фазовый спектр, являются незначительными.

Из-за вариаций спектра шума спектральное вычитание может привести к получению отрицательных оценок амплитудного спектра. Это особенно вероятно при понижении отношения сигнал/шум. Для того чтобы избежать появления отрицательных амплитудных значений, результат спектрального вычитания обрабатывается посредством использования функций отображения следующего вида:

$$T\left[\left|\hat{X}(\omega)\right|\right] = \begin{cases}\left|\hat{X}(\omega)\right|, & \left|\hat{X}(\omega)\right| > |Y(\omega)|,\\ \beta|Y(\omega)|, & \left|\hat{X}(\omega)\right| \le |Y(\omega)|,\end{cases} \tag{4.8}$$

где β – масштабирующий коэффициент.

Спектральное вычитание можно применять как к амплитудному спектру, так и к спектру мощности.

4.3 Применение метода спектрального взвешивания к нестационарным шумам

Использование методов параметрического моделирования и обработки с временным масштабированием позволяет выполнить более эффективную реализацию метода спектрального вычитания для нестационарных шумов, связанных со скоростью вращения механизма.

Предлагаемое решение состоит из следующих шагов (как показано на рисунке 4.3):

1) получение текущей скорости вращения при помощи внешнего источника либо оценки основного тона шума из зашумленного сигнала;

2) временное масштабирование сигнала с учетом полученной скорости вращения;

3) определение тональных компонент и их подавление путем узкополосной фильтрации;

4) определение голосовой активности и выделение фрагментов без полезного сигнала, оценка СПМ широкополосных компонент шума;

5) спектральное вычитание широкополосных компонент шума;

6) обратное временное масштабирование обработанного сигнала и возвращение его в линейный временной масштаб. Основные шаги обработки более детально описаны ниже.

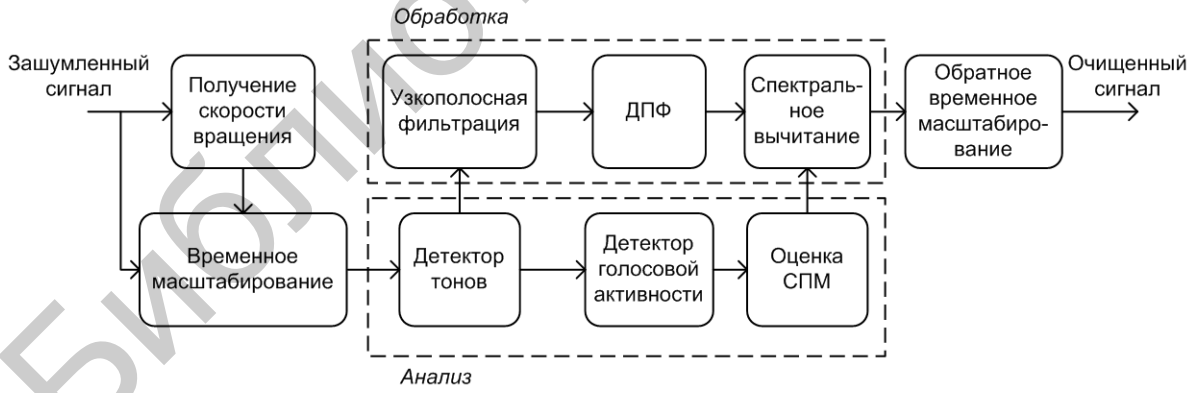


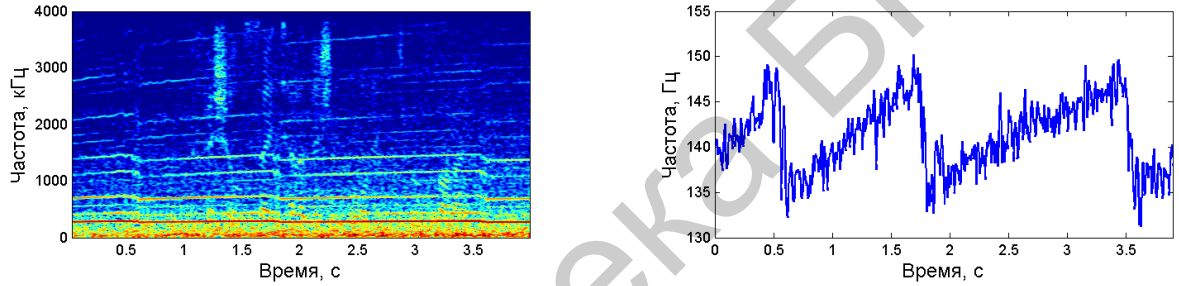
Рисунок 4.3 – Схема применения спектрального взвешивания с временным масштабированием для подавления нестационарных шумов

Оценка скорости вращения. При отсутствии возможности физически измерить скорость вращения при помощи внешних датчиков необходимо оценивать ее непосредственно из зашумленного сигнала.

Методы оценки контура мгновенной частоты основного тона подходят для этой задачи наиболее хорошо, поскольку позволяют оценивать постоянно

меняющуюся скорость вращения. Предлагается использовать алгоритмическую основу оценщика мгновенной частоты основного тона IRAPT, поскольку он позволяет использовать мгновенные синусоидальные параметры для оценки периодичности сигнала. Частота основного тона шума оценивается следующим образом.

Сигнал раскладывается на субполосные составляющие при помощи ДПФ-модулированного банка фильтров, затем субполосные составляющие описываются в виде мгновенных гармонических параметров и вычисляется функция генерации кандидатов основного тона, затем выполняется слежение за функцией генерации кандидатов в смежные моменты времени. Траектория изменения локальных максимумов определяется при помощи динамического программирования, которое накладывает ограничения на допустимую скорость изменения основного тона. Процедура определения скорости вращения при помощи данного подхода подразумевает алгоритмическую задержку в пределах 50 мс. Пример оценки контура мгновенной частоты шума показан на рисунке 4.4.



а б

а – зашумленный сигнал; б – полученный контур мгновенной частоты основного тона

Рисунок 4.4 – Оценка частоты основного тона шума из зашумленного сигнала

Спектральное вычитание в масштабированной временной области. Для того чтобы сделать частоту тональных компонент и СПМ шума более стабильными, выполняется процедура временного масштабирования, которая заключается в адаптивной дискретизации сигнала, согласованной с частотой основного тона шума. Результат временного масштабирования представлен на рисунке 4.5. Заметно, что амплитудный спектр шума в обработанном сигнале является менее изменчивым.

После выполнения временного масштабирования сигнал становится намного более удобным для выполнения шумоподавления методом спектрального взвешивания. Шумоподавление включает два основных шага: удаление тональных составляющих и удаление широкополосных составляющих. Обе операции используют кратковременный спектр Фурье, однако используются окна анализа различной длины.

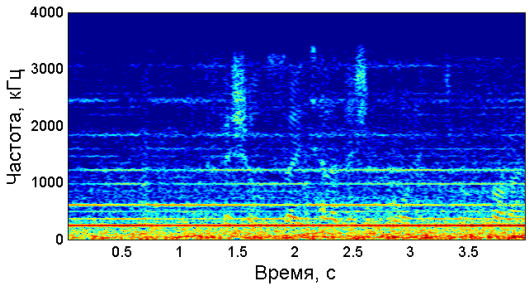
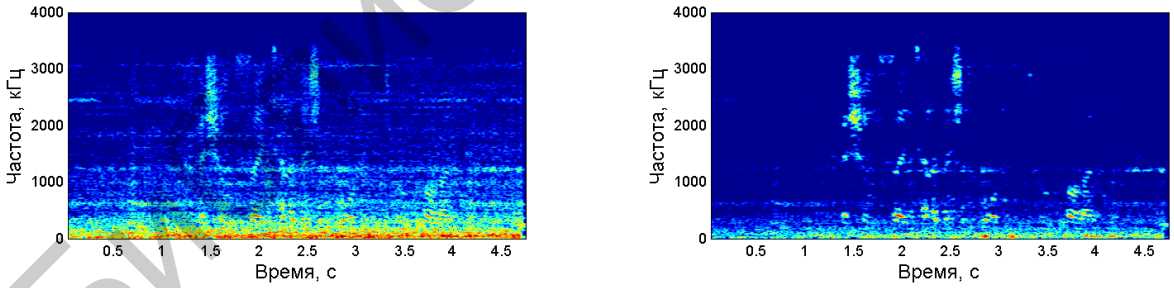


Рисунок 4.5 – Зашумленный сигнал после временного масштабирования с постоянной частотой основного тона шума

Для выявления и вычитания тональных составляющих используется длинное окно (около 100 мс). Учитывая, что тональные шумовые компоненты имеют постоянную частоту, они появляются в амплитудном спектре в виде острых пиков. Определитель шумовых тонов оценивает отсчеты амплитудного спектра и выявляет те из них, которые имеют большие значения по сравнению с соседними. Гармоники основного тона речи не определяются как шумовые тона, поскольку благодаря модуляциям частоты основного тона речи их энергия сглаживается на большом интервале наблюдения. Выявленные шумовые тона вычитаются из спектра. После вычитаются широкополосные шумовые составляющие, для чего используются короткие окна анализа (около 20 мс). Оценка СПМ широкополосных составляющих шума обновляется в местах, где отсутствует полезный сигнал. Для определения присутствия полезного сигнала используется определитель голосовой активности.

Учитывая, что благодаря временному масштабированию статистика шума стала менее зависима от скорости вращения, спектральное вычитание является весьма эффективным. На рисунке 4.6 показан результат подавления шума.



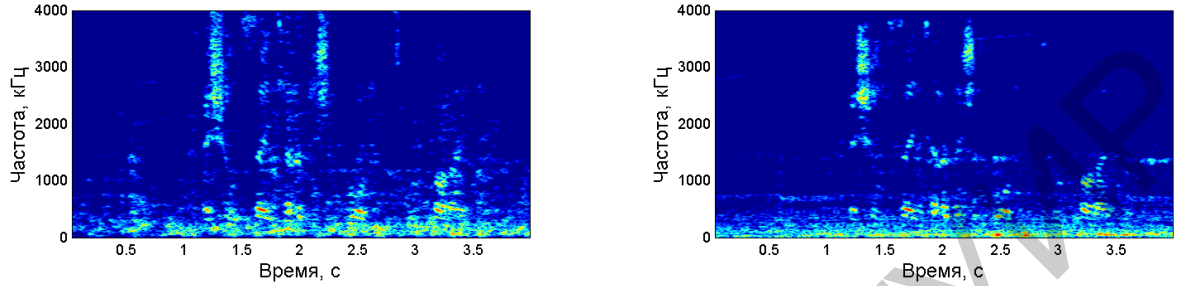
а – сигнал после вычитания тональных составляющих шума;
б – сигнал после вычитания широкополосных составляющих шума

Рисунок 4.6 – Результат спектрального вычитания, выполненного отдельно для тональных и широкополосных шумовых составляющих

После выполнения вычитания спектров узкополосных и широкополосных компонент реконструированный очищенный сигнал формируется путем обратного временного масштабирования.

Результаты экспериментов. Далее приводится практическое сравнение описанного метода спектрального вычитания для нестационарных шумов вращающихся механизмов с известным методом апостериорного анализа MAP.

На рисунке 4.7 показан результат обработки сигнала, зашумленного шумом движущегося автомобиля (шум записан в кабине болида Формулы-1) при помощи метода MAP и изложенного метода спектрального вычитания для нестационарных шумов.



а – результат шумоподавления методом MAP; б – результат шумоподавления спектральным вычитанием

Рисунок 4.7 – Сравнение метода MAP и метода спектрального вычитания для нестационарных шумов

Видно, что спектральное вычитание обеспечивает хорошее ослабление шума (18 дБ), в чем немного превосходит метод MAP (17 дБ). Кроме того, в полосе высоких частот (более 2 кГц) спектральное вычитание обеспечивает заметно меньший шумовой остаток. Гармоническая структура речевого сигнала также сохраняется лучше по сравнению с MAP. Однако из-за полностью автоматического определения тональных шумовых компонент некоторая часть (практически незаметная для восприятия) периодического шума сохранилась в сигнале после спектрального вычитания.

Субъективные оценки качества были получены путем прослушивания с использованием MOS-оценок. Оценивалось повышение качества сигнала из пар исходный/обработанный по шкале от минус 3 до плюс 3, где минус 3 соответствует существенному ухудшению качества, 0 – отсутствию изменений, плюс 3 – значительному повышению качества.

В прослушивании участвовала группа из 10 слушателей. В каждом тесте слушатель оценивал три пары речевых образцов и оценивал изменение качества. Общая средняя оценка для MAP составляет 1,62, в то время как для метода спектрального взвешивания – 1,71.

Необходимо также учитывать, что MAP использует внешний источник для оценки скорости вращения, в то время как в изложенном методе спектрального вычитания скорость вращения оценивалась непосредственно из зашумленного сигнала.

4.4 Обработка речевого сигнала в модуляционной области

Очистка речевого сигнала от сложных помех возможна путем фильтрации в модуляционной области. Понятия «модуляционного спектра» и «модуляционной области» определяются через понятие «модуляционной частоты». Если «акустическая частота» описывает частоту гармонических составляющих сигнала и используется для его декомпозиции при помощи преобразования Фурье, то «модуляционная частота» вводится как частота колебаний амплитудной огибающей и имеет другой физический смысл. Рассмотрим амплитудно-модулированный периодический сигнал с фиксированной частотой ω_c и ограниченным спектром:

$$s(t) = m(t)\cos(\omega_c t), \tag{4.9}$$

где модулирующий сигнал $m(t)$ принимает неотрицательные значения и ограничен по частоте таким образом, чтобы обеспечить возможность его восстановления из $s(t)$.

Тогда модуляционную частоту сигнала $s(t)$ можно получить, используя преобразование Фурье модулирующего сигнала:

$$M\left(e^{j\omega}\right) = F\{m(t)\} = \int_{-\infty}^{\infty} m(t)e^{j\omega t}dt. \tag{4.10}$$

Модель имеет более простую интерпретацию через узкополосный синусоидальный модулятор:

$$s(t) = \left(1 + \cos(\omega_m t)\right)\cos(\omega_c t). \tag{4.11}$$

Для восстановления модуляционного сигнал частота ω_m должна быть ниже несущей частоты ω_c. Для обеспечения неотрицательных значений модуляционного сигнала используется постоянное смещение. Без потери общности можно предположить, что модуляционный сигнал нормализован и принимает значения в диапазоне [–1; 1].

Степень влияния модуляционных частот хорошо исследована путем субъективной оценки разборчивости фильтрованной в модуляционной области речи. Эксперименты проводились раздельно для модуляционных фильтров низких и высоких частот, что позволило выделить наиболее важную полосу. В результате были сформулированы следующие основные особенности обработки речи модуляционной области:

1) амплитудный спектр речевого сигнала может быть ограничен сверху частотой 16 Гц без заметного снижения разборчивости для людей с нормальным слухом;

2) амплитудный спектр речевого сигнала может быть ограничен снизу частотой 4 Гц без заметного снижения разборчивости для людей с нормальным слухом;

3) слушатели лишь частично могут понимать речь в условиях идеальной акустической обстановки, если амплитудный спектр речевого сигнала ограничен сверху частотой 2 Гц либо снизу частотой 32 Гц; (при расширении полосы пропускания модуляционных фильтров разборчивость повышается);

4) согласные звуки деградируют сильнее гласных при ограничении модуляционного спектра.

Отметим, что основные модуляционные частоты, влияющие на разборчивость речи, находятся в диапазоне от 1 до 16 Гц с пиком около 3–5 Гц; кроме того, более 95 % модуляционных компонент речевого сигнала сконцентрированы в данном диапазоне. Это предположительно обусловлено количеством слогов, произносимых диктором за одну секунду.

4.5 Шумоподавление на основе фильтрации в модуляционной области

На основе приведенных выше положений были предложены два наиболее известных метода модуляционной фильтрации речи для подавления аддитивного шума и реверберации: фильтрация модуляционного спектра – RASTA и перцепционное линейное предсказание – PLP (Perceptual Linear Prediction).

Основная идея данных алгоритмов заключается в том, что модуляционные компоненты сигнала, не входящие в важный для восприятия диапазон, можно удалить без существенной потери разборчивости. В результате шумовые составляющие сигнала (стохастические, полигармонические помехи и реверберация) будут подавлены, поскольку существенная часть их энергии находится за пределами речевой полосы модуляционного спектра.

Алгоритм обработки речи на основе RASTA-PLP состоит из следующих основных шагов:

1) вычисление энергии сигнала в критических полосах;

2) компрессия амплитудных значений путем статической нелинейной трансформации;

3) фильтрация амплитудных огибающих каждого из субполосных сигналов;

4) растяжение амплитудного спектра путем обратной статической нелинейной трансформации;

5) умножение амплитудных значений на кривую равной громкости и возведение в степень 0,33 для выравнивания громкости.

Для нелинейного преобразования амплитуд (компрессии/растяжения) обычно используются логарифмические либо степенные функции. Логарифмическая трансформация приводит к тому, что такие искажения, которые представляют собой свертку во временной области (например, реверберация), проявляются как аддитивный шум в логарифмическом амплитудном спектре. Од-

нако в случае зашумления речи некоррелированным аддитивным сигналом возникают трудности, поскольку такой шум аддитивен в линейном спектре, но в логарифмическом становится зависим от самого сигнала, что делает невозможным его удаление путем фильтрации по частотным полосам. Решение данной проблемы заключается в использовании следующей функции преобразования:

$$Y(k, n\Delta t) = \ln\big(1 + JX(k, n\Delta t)\big), \tag{4.12}$$

где $X(k, n\Delta t)$ – амплитуда сигнала в k-м канале;

J – положительная константа, зависящая от типа сигнала.

Такое преобразование линейно для малых значений амплитуды и является логарифмическим для больших величин. Формула обратного преобразования имеет следующий вид:

$$Y(k, n\Delta t) = \frac{e^{Y(k, n\Delta t)} - 1}{J}. \tag{4.13}$$

4.6 Синтез модуляционного фильтра

В качестве фильтра, используемого для обработки сигнала в модуляционной области, авторами RASTA предложен БИХ-фильтр с передаточной функцией:

$$H(z) = 0{,}1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0{,}98z^{-1}}. \tag{4.14}$$

Нижняя граница полосы пропускания фильтра определяет максимальную скорость изменения логарифмического спектра, который отбрасывается в процессе фильтрации, в то время как верхняя граница полосы пропускания определяет максимальную скорость изменения сохраняемого спектра. На рисунке 4.8 показана амплитудно-частотная характеристика модуляционного полосового фильтра RASTA. Как следует из представленных результатов, начиная с частоты 12 Гц коэффициент усиления уменьшается со скоростью 6 дБ на октаву.

Основным недостатком алгоритмов на основе RASTA является то, что сигнал фильтруется без учета характера и степени зашумления. Поэтому системы на основе данных алгоритмов ориентированы на конкретную акустическую среду и не могут подстраиваться к изменяющимся условиям шумовой обстановки. Более того, чистая речь после обработки RASTA-системой звучит неприятно из-за наличия музыкальных тонов.

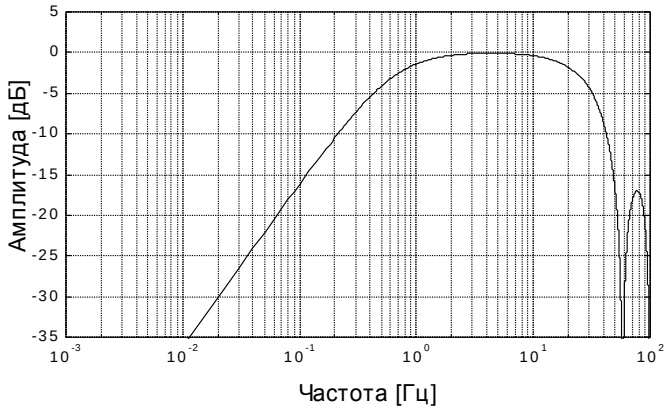


Рисунок 4.8 – Амплитудно-частотная характеристика модуляционного полосового фильтра RASTA

В некоторых исследованиях используются полосовые КИХ-фильтры (полоса пропускания 1–16 Гц) с различными амплитудно-частотными характеристиками. Известно применение модуляционного фильтра, который с целью улучшения разборчивости речи усиливал модуляционные компоненты в диапазоне 2–8 Гц. Однако оба подхода не позволяют избавиться от вышеупомянутого недостатка. В более поздних исследованиях предпринимается попытка выбора модуляционного фильтра с учетом оценки условий окружающей акустической среды и для повышения качества речи вместо одного модуляционного полосового фильтра с постоянной характеристикой используется их набор.

Поскольку существует прямая зависимость между слоговой разборчивостью речи и энергией в полосе определенных частот модуляции, целесообразно выполнять слежение во времени за характером изменения модуляционного спектра и соответствующим образом менять полосу пропускания модуляционного фильтра, т. е. синтезировать модуляционный фильтр с изменяющимися во времени параметрами. Можно в качестве прототипа использовать параметрический перестраиваемый фильтр, у которого в каждом частотном канале *k* во времени изменяется коэффициент a_{0k}, определяющий полосу пропускания фильтра $\Delta\omega_k$. Центральная частота ω_0 сохраняется постоянной. Передаточная функция фильтра имеет вид

$$H_k(z) = a_{0k}\frac{1 - z^{-2}}{1 + (a_{0k} - 1)gz^{-1} + (1 - 2a_{0k})z^{-2}}, \tag{4.15}$$

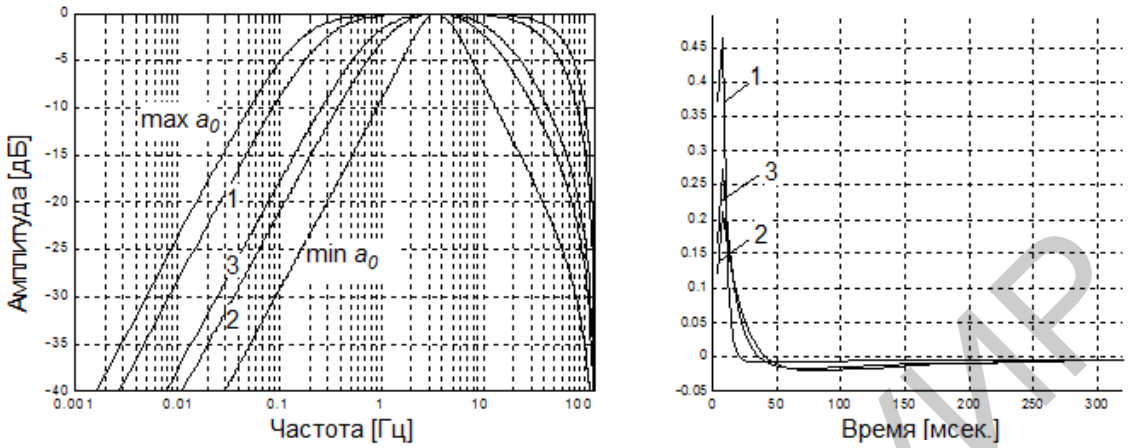
где $a_{0k} = \frac{\Delta\omega_k \Delta t}{2 + \Delta\omega_k \Delta t}$;

$g = 2\cos(\omega_0 \Delta t)$;

$\Delta t = \frac{M}{f_s}$, M – число полос банка фильтров;

f_s – частота дискретизации.

На рисунке 4.9 изображены импульсные и частотные характеристики данного фильтра для разных значений коэффициента a_{0k}.



а б

а – амплитудно-частотные характеристики; б – импульсные характеристики

Рисунок 4.9 – Перестраиваемый фильтр для обработки речевого сигнала в модуляционной области

4.7 Подавление шумов путем фильтрации в модуляционной области

Для субполосной декомпозиции речевого сигнала обычно используют банк фильтров. Фильтрация выполняется в фиксированных полосах, которые заранее определены на этапе расчета банка фильтров. Упрощенно схема обработки сигнала в модуляционной области может быть представлена, как на рисунке 4.10.

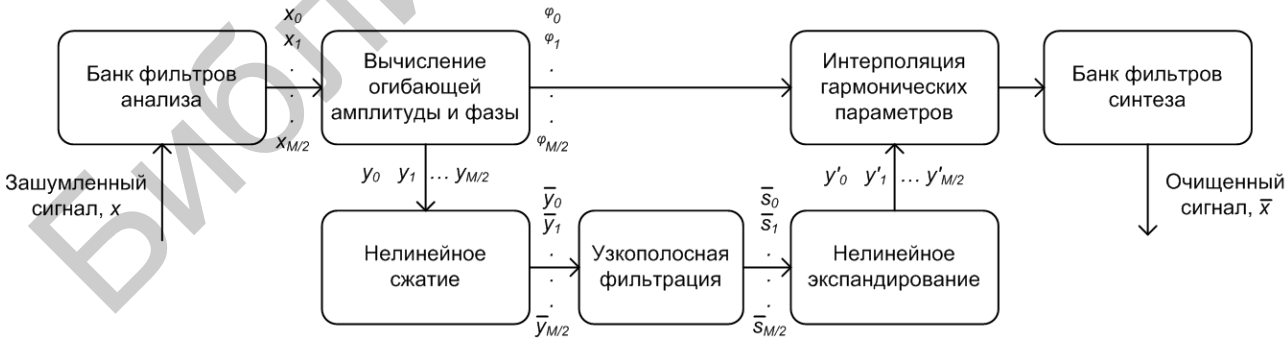


Рисунок 4.10 – Схема обработки сигнала в модуляционной области для подавления шумов

Согласно приведенной схеме алгоритм состоит из следующих шагов.

Шаг 1. Речевой сигнал $x(n/f_s)$ разделяется на M частотных полос со следующими нормализованными центральными частотами: $(2k+1)\pi/2M$, причем каждый канал характеризуется своей импульсной характеристикой $p_k(n)$, а

также частотной характеристикой $P_k(f)$ для $0 \le k \le M/2$. Ширина каждой полосы не должна превышать частоту основного тона сигнала для того, чтобы исключить попадания нескольких гармоник основного тона в один канал.

Шаг 2. В каждом канале $k(0 \le k \le M/2)$ вычисляется амплитудная огибающая сигнала $x_k(nM/f_s)$:

$$y_k\left(\frac{nM}{f_s}\right) = \sqrt{\mathrm{Re}^2[x_k(nM/f_s)] + \mathrm{Im}^2[x_k(nM/f_s)]} \tag{4.16}$$

и фаза

$$\varphi_k\left(\frac{nM}{f_s}\right) = -\mathrm{arctg}\frac{\mathrm{Im}[x_k(nM/f_s)]}{\mathrm{Re}[x_k(nM/f_s)]}. \tag{4.17}$$

Затем выполняется трансформация амплитуды огибающей спектра $y_k(nM/f_s)$ путем нелинейного статического сжатия $\bar{y}_k\left(\frac{nM}{f_s}\right) = \ln\left(1 + +1000\left|y_k\left(\frac{nM}{f_s}\right)\right|\right)$ и фильтрация огибающей амплитуды спектра $\bar{y}_k\left(\frac{nM}{f_s}\right)$ модуляционным фильтром. Параметр a_{0k}, определяющий полосу пропускания модуляционного фильтра в каждом канале, выбирается согласно зашумлению сигнала. Далее выполняется трансформация огибающей амплитудного спектра $\bar{s}_k(nM/f_s)$ в линейный масштаб путем обратного нелинейного преобразования

$$y'_k = \left(\frac{nM}{f_s}\right) = \frac{e^{\bar{y}_k\left(\frac{nM}{f_s}\right)-1}}{1000}.$$

Шаг 3. Восстановление речевого сигнала $\bar{x}_k(n/f_s)$ при помощи банка фильтров синтеза.

4.8 Использование мгновенных синусоидальных параметров для повышения качества фильтрации в модуляционной области

В приведенном выше алгоритме частотные полосы равномерно перекрывают весь диапазон сигнала и не пересекаются. Причем, с одной стороны, число полос должно быть максимально большим, чтобы увеличить частотное разрешение алгоритма обработки, с другой стороны, каждая полоса должна быть широкой для сохранения эффективного диапазона модуляционного спектра. Если частота дискретизации составляет 8 кГц, то *M* следует выбирать не меньше 57, но и не более 130. Данные ограничения приводят к неточному разделению обрабатываемого сигнала на речь и шум, что выражается в подавлении полезного сигнала и повышенном уровне остаточного шума, а также излишнего «отбеливания» речи.

Таким образом, целесообразно использование иного способа расчета амплитудной огибающей, основанного на суммировании мгновенных амплитуд

нескольких каналов согласно текущему значению мгновенной частоты, как схематично показано на рисунке 4.11.

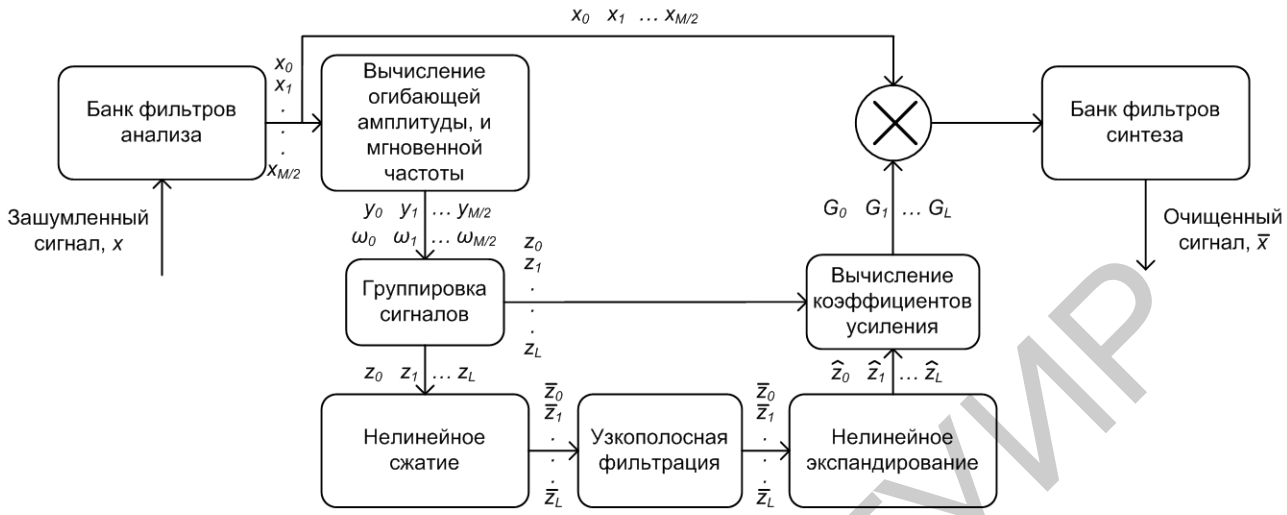


Рисунок. 4.11 – Схема обработки сигнала в модуляционной области для подавления шумов с использованием мгновенной частоты

Ниже приводится описание алгоритма.

Шаг 1. Речевой сигнал $x(n/f_s)$ разделяется на M частотных полос со следующими нормализованными центральными частотами: $(2k+1)\pi/2M$. Ширина каждой полосы фиксирована (более 30 Гц) и не зависит от числа каналов. M выбирается сколь угодно большим в зависимости от производительности вычислительной платформы (в экспериментальной реализации использовалось 2048 каналов).

Шаг 2. В каждом канале $k(0 \le k \le M/2)$ амплитудная огибающая $y_k(nM/f_s)$, фаза $\varphi_k(nM/f_s)$ и мгновенная частота вычисляются по формулам (4.16) и (4.17).

Шаг 3. Исходя из полученных значений мгновенной частоты $\omega_k(nM/f_s)$, амплитудные огибающие группируются в L каналов, причем $L < M$ (в экспериментальной реализации использовалось $L = 64$). Результирующие огибающие вычисляются как среднее квадратическое от мгновенных амплитудных значений субканальных сигналов, входящих в одну группу:

$$z_l\left(\frac{nM}{f_s}\right) = \sqrt{\frac{\sum_{i\in\text{группа}_l} y_i\left(\frac{nM}{f_s}\right)^2}{M/L}}, \quad 1 \le l \le L. \tag{4.18}$$

Выполняется трансформация амплитуды огибающей спектра $z_l(nM/f_s)$ путем нелинейного статического сжатия $\bar{z}_l\left(\frac{nM}{f_s}\right) = \ln\left(1 + 1000\left|z_l\left(\frac{nM}{f_s}\right)\right|\right)$. Выполняется фильтрация огибающей амплитуды спектра $\bar{z}_l(nM/f_s)$ модуляцион-

ным фильтром. Параметр a_{0k} выбирается согласно процедуре, описанной в [11]. Затем выполняется перевод амплитуды огибающей спектра $\bar{z}_l(nM/f_s)$ в линейный масштаб путем обратного нелинейного преобразования:

$$\hat{z}_l\left(\frac{nM}{f_s}\right) = \frac{e^{\bar{z}_l\left(\frac{nM}{f_s}\right)-1}}{1000}. \tag{4.19}$$

Шаг 4. Вычисление коэффициентов усиления G для каждого из L каналов:

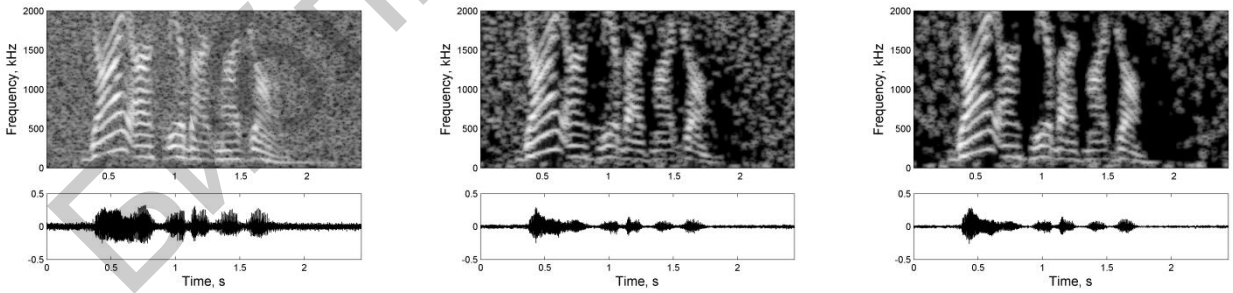
$$G_l = \frac{\hat{z}_l\left(\frac{nM}{f_s}\right)}{z_l\left(\frac{nM}{f_s}\right)}. \tag{4.20}$$

Амплитуда каждого из M исходных каналов умножается на соответствующий коэффициент G_l и выполняется восстановление речевого сигнала при помощи банка фильтров синтеза.

4.9 Результаты экспериментов

Сравнение полученного алгоритма на основе мгновенных синусоидальных параметров (далее МФ2) с известным [11] (далее МФ1) выполнено путем сравнения результатов обработки специально подготовленных речевых сигналов, зашумленных помехами различных типов. Для сравнения использовались объективные оценки на основе анализа спектра сигналов и при помощи субъективных оценок на основе результатов прослушивания.

На рисунке 4.12 представлен результат обработки сигнала, зашумленного белым шумом.



а б в

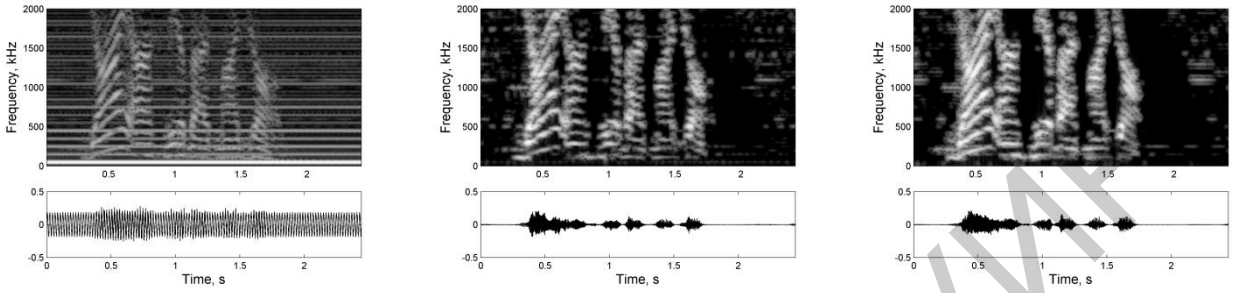
а – исходный речевой сигнал; б – МФ1; в – МФ2

Рисунок 4.12 – Результат обработки сигнала, зашумленного белым шумом

В обработанных сигналах энергия аддитивного шума снижена на 8 дБ (МФ1) и на 10 дБ (МФ2). Видно, что оба обработанных сигнала содержат му-

зыкальные тона, однако интенсивность тонов заметно ниже для алгоритма МФ2.

На рисунке 4.13 представлен результат обработки сигнала, зашумленного сетевой помехой. Сетевая помеха представляет собой совокупность гармонических компонент с медленно изменяющимися амплитудами и поэтому эффективно задерживается модуляционными фильтрами.



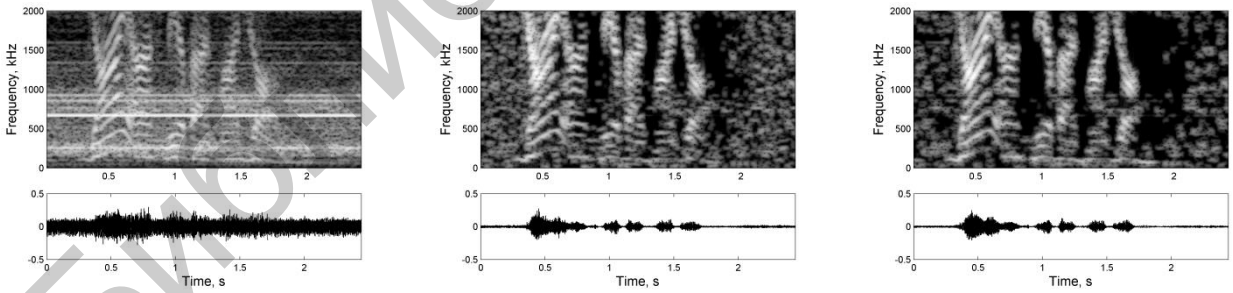
а б в

а – исходный речевой сигнал; б – МФ1; в – МФ2

Рисунок 4.13 – Результат обработки сигнала, зашумленного сетевой помехой

На участках, где отсутствует голос диктора, снижение энергии сетевой помехи составляет примерно 33 и 40 дБ для алгоритмов МФ1 и МФ2 соответственно.

На рисунке 4.14 представлен результат обработки сигнала, зашумленного шумом пылесоса, который состоит из периодических и непериодических составляющих. Особенностью данной помехи является ее нестационарность.



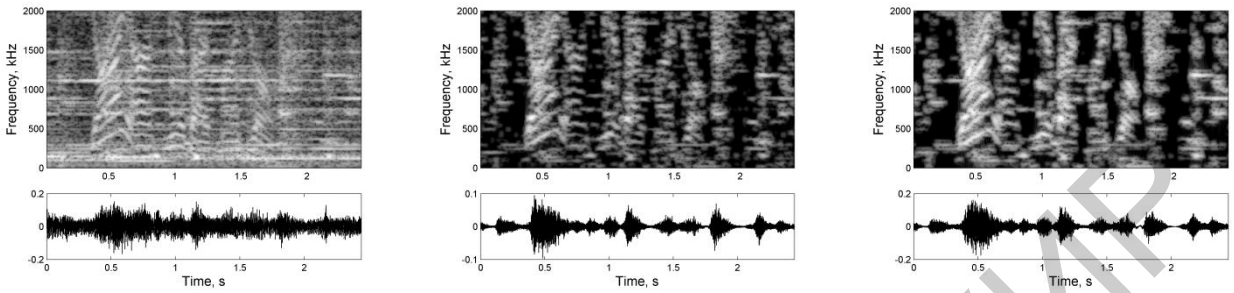
а б в

а – исходный речевой сигнал; б – МФ1; в – МФ2

Рисунок 4.14 – Результат обработки сигнала, зашумленного шумом пылесоса

На участках, где отсутствует голос диктора, энергия помехи в среднем снижена на 20 и 23 дБ для алгоритмов МФ1 и МФ2 соответственно.

На рисунке 4.15 представлен результат обработки речи на фоне музыки. Автоматическое отделение речи от музыки в одноканальных системах представляет собой особенно сложную задачу. Тем не менее использование модуляционных фильтров обеспечивает некоторое подавление тональных и переходных составляющих, характерных для музыкальных инструментов.



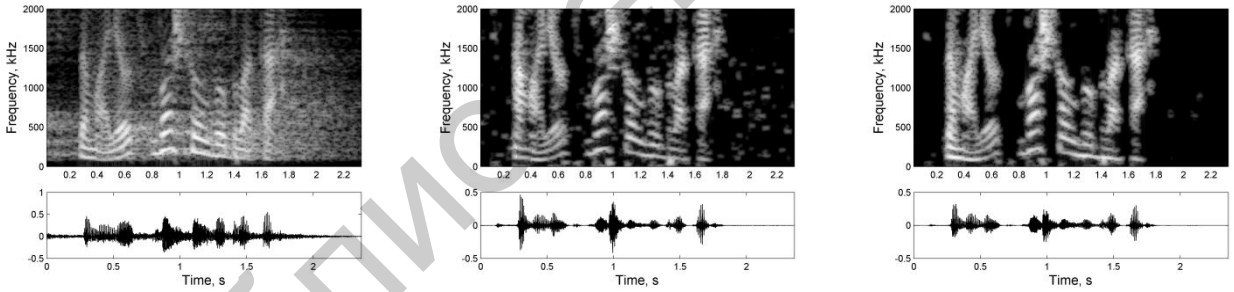
а б в

а – исходный речевой сигнал; б – МФ1; в – МФ2

Рисунок 4.15 – Результат обработки сигнала, зашумленного фоновой музыкой

Снижение энергии музыки составляет примерно 4 дБ для обоих алгоритмов модуляционной фильтрации.

На рисунке 4.16 представлен результат обработки речи, зашумленной реверберацией.



а б в

а – исходный речевой сигнал; б – МФ1; в – МФ2

Рисунок 4.16 – Результат обработки сигнала, зашумленного реверберацией

Во всех приведенных ранее примерах предложенный алгоритм фильтрации обеспечивает более низкий уровень музыкальных тонов и других артефактов, более точно сохраняет форму спектральной огибающей и гармоническую структуру речевого сигнала. Различия особенно заметны в низкочастотной области спектра (ниже 500 Гц).

Результаты субъективной оценки качества очистки речевых сигналов приведены в таблице 4.1. В процессе эксперимента использовались записи мужского голоса продолжительностью 3 мин, зашумленные помехами различ-

ных типов. Четырем слушателям была предложена субъективная оценка (по пятибалльной шкале) качества подавления шума в обработанных сигналах по следующим показателям:

1) разборчивость (точность восприятия речевого сообщения: 5 – полная, 4 – почти полная, 3 – неполная, 2 – частичная, 1 – отсутствует);

2) субъективное зашумление (степень зашумления: 5 – высокий уровень зашумления, ощущается сильный дискомфорт при прослушивании, 4 – умеренный дискомфорт при прослушивании, 3 – сохраняется средняя комфортность прослушивания, 2 – сохраняется высокая комфортность прослушивания, 1 – нет зашумления);

3) качество реконструкции (естественность звучания обработанного речевого сигнала: 5 – абсолютно натуральное звучание, 4 – почти натуральное звучание с небольшим уровнем артефактов, 3 – средний уровень артефактов, 2 – высокий уровень артефактов, 1 – речь полностью неестественная).

Полученные результаты свидетельствуют о том, что предложенный алгоритм фильтрации речи в модуляционной области обеспечивает более высокое субъективное качество шумоподавления. Это достигается как за счет снижения уровня слышимых артефактов, так и за счет более точного восстановления спектральных огибающих речевого сигнала.

Таблица 4.1 – Результаты субъективной оценки качества очистки речевых сигналов

Тип помехи	Разборчивость			Субъективное зашумление			Качество реконструкции		
	Исх.	МФ1	МФ2	Исх.	МФ1	МФ2	Исх.	МФ1	МФ2
Белый шум	4,0	4,25	4,25	4	3	2,75	–	2,75	3
Сетевая помеха	3,25	4,0	4,25	4,5	2,5	2,25	–	3,5	3,75
Шум пылесоса	2,75	3,5	3,75	4,75	3	2,5	–	3	3,25
Фоновая музыка	2	2,5	2,75	4,5	4,25	4,25	–	3,25	3,25
Реверберация	4	4,25	4,25	3	1,5	1,25	–	3,5	3,75

4.10 Выводы

Рассмотрен метод шумоподавления на основе спектрального вычитания, который в отличие от существующих применим для подавления нестационарных шумов, создаваемых вращающимися механизмами. Метод использует временное масштабирование сигнала для снижения влияния изменения скорости вращения на статистику шума. Учитывая, что мгновенная скорость вращения пропорциональна частоте основного тона шума, ее во многих случаях можно выделить непосредственно из зашумленного сигнала при помощи специального алгоритма слежения. Показано, что данный метод подавления шума является эффективным и предпочтительным в сравнении с существующим методом шу-

моподавления MAP, поскольку позволяет выполнять очистку сигнала от широкополосных шумовых составляющих согласованно со скоростью вращения.

Рассмотрен способ очистки речевого сигнала от шума, основанный на фильтрации в модуляционной области RASTA (RelAtive SpecTrA). Способ является универсальным и позволяет подавлять широкий класс акустических помех. Исследование информационной значимости модуляционных частот речи является одним из приоритетных современных направлений развития систем шумоподавления и распознавания речи в условиях агрессивной акустической обстановки. Высокий потенциал подхода модуляционной фильтрации для повышения разборчивости речи в шумах объясняется его физиологической мотивацией и тесной связью с психоакустикой. Практическим достоинством подхода является универсальность и автоматическая адаптивность к помехам разных типов.

Предложен способ фильтрации в модуляционной области на основе мгновенных параметров синусоидальной модели. На основании результатов практических экспериментов показано, что предложенный способ обеспечивает более высокий коэффициент ослабления шума и меньшую степень деградации полезного речевого сигнала в сравнении с известным алгоритмом. Полученные результаты свидетельствуют о применимости предложенного алгоритма для обработки речевых сигналов, зарегистрированных в различных акустических условиях. Основным достоинством алгоритма является более низкий уровень слышимых артефактов и более высокое качество реконструкции речевого сигнала. Наиболее хорошо алгоритм применим для очистки речевого сигнала от тональных шумов высокой интенсивности (таких как сетевая помеха) и реверберации.

ЛИТЕРАТУРА

1 Оппенгейм, А. Цифровая обработка сигналов / А. Оппенгейм, Р. Шафер. – М. : Техносфера, 2006. – 853 с.

2 Лайонс, Р. Цифровая обработка сигналов / Р. Лайонс ; пер. с англ. – 2-е изд. – М. : ООО «Бином-Пресс», 2006. – 656 с.

3 Рабинер, Л. Теория и применение цифровой обработки сигналов / Л. Рабинер, Б. Гоулд. – М. : Мир, 1978. – 848 с.

4 Сергиенко, А. Б. Цифровая обработка сигналов / А. Б. Сергиенко. – СПб. : Питер, 2002. – 608 с.

5 Солонина, А. И. Алгоритмы и процессоры цифровой обработки сигналов / А. И. Солонина, Д. А. Улахович, Л. А. Яковлев. – СПб. : БХВ-Петербург, 2002. – 464 с.

6 Вашкевич, М. И. Косинусно-модулированные банки фильтров с фазовым преобразованием: реализация и применение в слуховых аппаратах / М. И. Вашкевич, И. С. Азаров, А. А. Петровский – М. : Горячая линия – Телеком, 2014. – 210 с.

7 Гольденберг, Л. М. Цифровая обработка сигналов : учеб. пособие для вузов / Л. М. Гольденберг, В. Д. Матюшкин, М. Н. Поляк. – М. : Радио и связь, 1990. – 315 с.

8 Каппелини, В. Цифровые фильтры и их применение / В. Каппелини, А. Константинидис, П. Эмилиани. – М. : Радио и связь, 1983. – 350 с.

9 Хемминг, Р. В. Цифровые фильтры / Р. В. Хемминг. – М. : Сов. радио, 1980. – 224 с.

10 Голд, Б. Цифровая обработка сигналов / Б. Голд, Ч. Рэйдер. – М. : Сов. радио, 1973. – 368 с.

11 Bashun, J. Speech enhancement for cochlear implants based on the reducing slow temporal modulations / J. Bashun, A. Petrovsky // Proc. of The Acoustic Congress, Rom, Italy. – 2000.

Св. план 2015, поз. 23

Учебное издание

Петровский Александр Александрович
Вашкевич Максим Иосифович
Азаров Илья Сергеевич

ЦИФРОВАЯ ОБРАБОТКА АУДИО- И ВИДЕОДАННЫХ

ПОСОБИЕ

Редактор *Е. И. Герман*
Корректор *Е. Н. Батурчик*
Компьютерная правка, оригинал-макет *В. М. Задоля*

Подписано в печать 04.05.2017. Формат 60x84 1/16. Бумага офсетная. Гарнитура «Таймс».
Отпечатано на ризографе. Усл. печ. л. 3,84. Уч.-изд. л. 4,0. Тираж 70 экз. Заказ 276.

Издатель и полиграфическое исполнение: учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий №1/238 от 24.03.2014,
№2/113 от 07.04.2014, №3/615 от 07.04.2014.
ЛП №02330/264 от 14.04.2014.
220013, Минск, П. Бровки, 6