УДК 621.371.39:681.322.01

ГАРМОНИЧЕСКАЯ МОДЕЛЬ РЕЧЕВОГО СИГНАЛА: ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ И ИХ КВАНТОВАНИЕ

А.Н. ПАВЛОВЕЦ[1], П. ЗУБРЫЦКИ[2], А.А. ПЕТРОВСКИЙ[1]

[1]*Белорусский государственный университет информатики и радиоэлектроники*
П. Бровки, 6, Минск, 220013, Беларусь

[2]*Технический университет, Белосток, Польша*

Поступила в редакцию 29 мая 2007

Рассматривается метод определения параметров гармонической модели речевого сигнала с последующим их квантованием. Особенностью метода является применение цикла с обратной связью для определения гармонических амплитуд и фаз. Предлагается также использование закономерностей психоакустики в процедуре квантования амплитуд.

Ключевые слова: гибридный вокодер, гармоническая модель, частота основного тона, квантование векторов переменной длины, психоакустика.

Введение

Большинство сигналов в природе, включая речь и музыку, могут быть описаны при помощи гармонической модели, которая определяется следующим набором параметров: фундаментальной частотой, амплитудой и фазой каждой частотной компоненты. Гармонический сигнал генерируется серией синусоид или гармонических компонент, частоты которых являются целочисленным кратным некоторой фундаментальной частоты. Данная модель является весьма эффективным решением для большого количества приложений кодирования сигнала, так как позволяет представить сигнал с помощью достаточно компактного набора параметров.

Первые попытки представления речевого сигнала с помощью гармонической модели датируются началом 1980-х гг. [1]. В дальнейшем в системах анализа–синтеза речи данное представление стало уточняться и дополняться описанием сигнала-остатка в форме шумовой модели [2], что позволяет повысить точность представления речевого сигнала, а вместе с тем и качество.

Некоторые сегменты речевого сигнала сложно разделить на периодическую и апериодическую составляющие, используя гармоническую и шумовую модели. Это происходит при попадании в сегмент взрывных звуков, наличии в сегменте одновременно гласных и глухих согласных, присутствии каких-то локальных явлений. Следующей ступенью развития представления речевого сигнала стала гибридная модель [3], предусматривающая три возможных класса для сегмента речи — вокализованный, невокализованный, переходный. Особенностью ее является анализ–синтез переходных сегментов во временной области, в то время как вокализованные и невокализованные сегменты обрабатываются в частотной области.

Модель речевого сигнала, рассматриваемая в данной работе, предполагает классификацию речевого сегмента с точки зрения возможности декомпозиции его на

гармоническую и шумовую компоненты. Признаком такой возможности является вокализованность сегмента (рис. 1).

В такой модели важнейшим аспектом является корректное определение гармонической и шумовой компонент. Существуют различные подходы к их разделению. Так, например, в [2] сегмент речи представлен суммой гармонической и шумовой составляющих, спектры которых находятся соответственно до и после так называемой максимальной частоты вокализованности. Подход [4] характеризуется определением вокализованности в определенных частотных полосах.

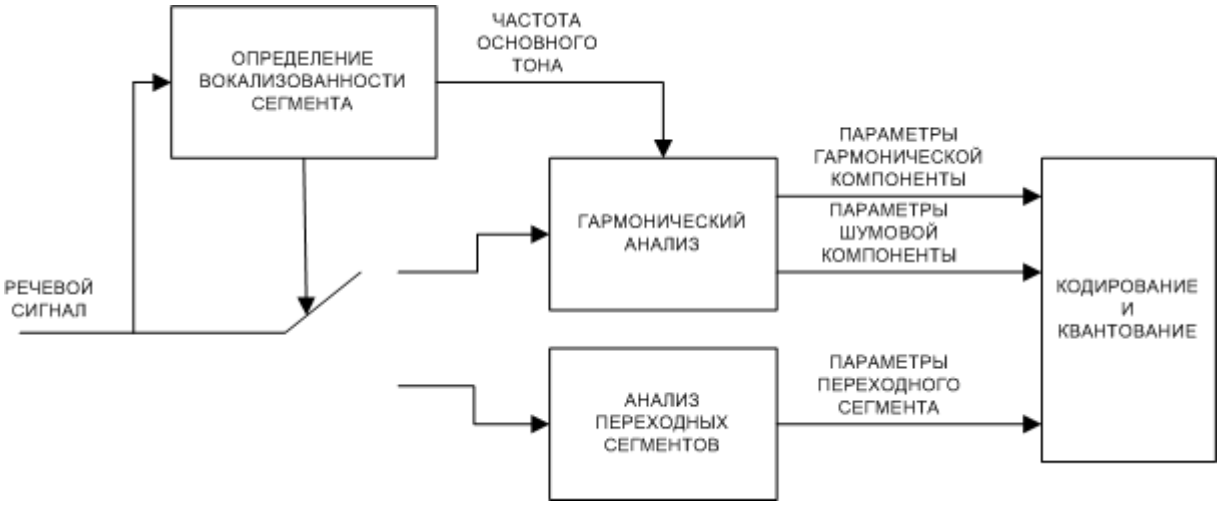


Рис. 1. Схема декомпозиции речевого сигнала в вокодере, основанном на раздельном кодировании гармонической, шумовой и переходной компонент

Данные подходы не вполне адекватно описывают речевой сигнал, поскольку в них используются бинарные решения о вокализованности сигнала в целой полосе частот. Отличие рассматриваемой модели анализа–синтеза речи от вышеупомянутых состоит в использовании для декомпозиции речевого сигнала дискретного преобразования Фурье, согласованного с изменением контура частоты основного тона [5–7]. Этот подход позволяет разделить гармоническую и шумовую составляющие во всем речевом спектре.

Одним из фундаментальных вопросов в приложениях кодирования на базе гармонических моделей является квантование *гармонических амплитуд*, так как качество реконструированной речи в параметрических вокодерах в большой степени зависит от качества квантования параметров гармонической компоненты, несущей основную информацию о кодируемом речевом сигнале.

В настоящее время известно достаточно большое количество подходов кодирования последовательности гармонических амплитуд. Скалярное квантование, например, квантует каждый элемент индивидуально; тем не менее векторное квантование [8] является более предпочтительным подходом для современных алгоритмов низкоскоростных кодеров речи, что обусловлено улучшенным качеством последнего. Традиционные векторные квантователи строятся с учетом фиксированной длины векторов. В последних работах удалось добиться достаточно высокого качества квантования гармонических амплитуд благодаря применению схемы расщепленного векторного квантования линейных спектральных пар, при этом прозрачное кодирование достигалось при скорости 23 бит/вектор [9]. Однако построение векторного квантователя с переменной длиной кодируемого вектора гармонических амплитуд выглядит более естественным решением ввиду того, что при этом не требуется осуществления дополнительных преобразований над входным вектором.

Таким образом, целью данной работы является разработка метода определения параметров гармонической модели и их квантования.

Определение параметров гармонической модели речи

Согласно модели (рис. 1), вокализованный сегмент речевого сигнала может быть представлен в виде суммы гармонической и шумовой составляющих:

$s(i)=h(i)+r(i)$. (1)

Гармоническая модель для описания речи впервые была предложена в [1]:

$$h(i)=\sum_{k=1}^{M} A_k \cos\left(2\pi k \frac{F_0}{F_s} i+\theta_k\right), \quad (2)$$

где F_0 — частота основного тона; A_k — амплитуда k-й гармонической компоненты; θ_k — фаза k-й гармонической компоненты; M — количество гармоник; F_s — частота дискретизации.

Данная модель, однако, не учитывает изменение частоты основного тона во времени. При достаточно малой величине окна анализа (до 25 мс) можно предположить, что это изменение носит линейный характер. Уточненная модель выглядит следующим образом:

$$h(i)=\sum_{k=1}^{M} A_k \cos\left(\frac{2\pi k i}{F_s}\left(F_0+\frac{\Delta F_0 i}{2N}\right)+\theta_k\right), \quad (3)$$

где ΔF_0 – изменение частоты основного тона за N отсчетов.

Следующим уточнением гармонической модели речи будет введение в формулу (3) фактора, учитывающего изменение (нарастание либо затухание) гармонических амплитуд с течением времени [10]:

$$h(i)=\sum_{k=1}^{M} A_k e^{-\beta i} \cos\left(\frac{2\pi k i}{F_s}\left(F_0+\frac{\Delta F_0 i}{2N}\right)+\theta_k\right), \quad (4)$$

где β — фактор изменения; предполагается, что значения гармонических амплитуд эволюционируют по экспоненциальному закону. Использование формулы (4) позволяет повысить точность представления нестационарных вокализованных речевых сегментов.

Для определения параметров гармонической модели речи (амплитуд и фаз) на заданном сегменте удобно проводить анализ, синхронизированный с изменением контура частоты основного тона (следящий анализ) [6]:

$$H_n(k)=\sum_{i=0}^{L-1} s_n(i)\ \exp\left(j\frac{2\pi k i}{F_s}\left(F_0+\frac{\Delta F_0 i}{2N}\right)\right) w_n(i), \quad j=\sqrt{-1}, \quad (5)$$

где $w_n(i)$ — временное окно; n — номер сегмента.Тогда амплитуды и фазы k-й гармоники определяются соответственно как

$$A_n(k)=\frac{\sqrt{\mathrm{Re}^2(H_n(k))+\mathrm{Im}^2(H_n(k))}}{\sum_{i=0}^{L-1} w(i)}, \quad (6)$$

$$\theta_n(k)=-\mathrm{arctg}\frac{\mathrm{Im}(H_n(k))}{\mathrm{Re}(H_n(k))} \quad (7)$$

Величину фактора изменения амплитуд β можно определить следующим образом [9]:

$$\beta=\ln(x)/T_0,\ x=\frac{S_0{}^T S_{T_0}}{S_{T_0}{}^T S_{T_0}}, \quad (8)$$

где $S_0 = [|s_0|,...,|s_{N-T_0-1}|]^T$, $S_{T_0} = [|s_{T_0}|,...,|s_{N-1}|]^T$; T_0 — значение периода основного тона в отсчетах, s_0, s_1,…, s_{N-1} — отсчеты речевого сигнала, N — длина сегмента.

Шумовая компонента выделяется путем вычитания из исходного сигнала синтезированной, согласно (4), гармонической компоненты:

$$r(i)=s(i)-h(i). \quad (9)$$

Поскольку гармоническая и шумовая модели кардинально отличаются друг от друга, важным аспектом является точная сепарация гармонической и шумовой компонент речи. Показателем точности в данном случае может служить отношение "гармоники / шум":

$$HNR = 10\lg\frac{E_h}{E_r}, \quad (10)$$

где E_h и E_r — энергии гармонической и шумовой компоненты соответственно.

Рассмотрим спектрограмму некоторого речевого фрагмента (рис. 2).

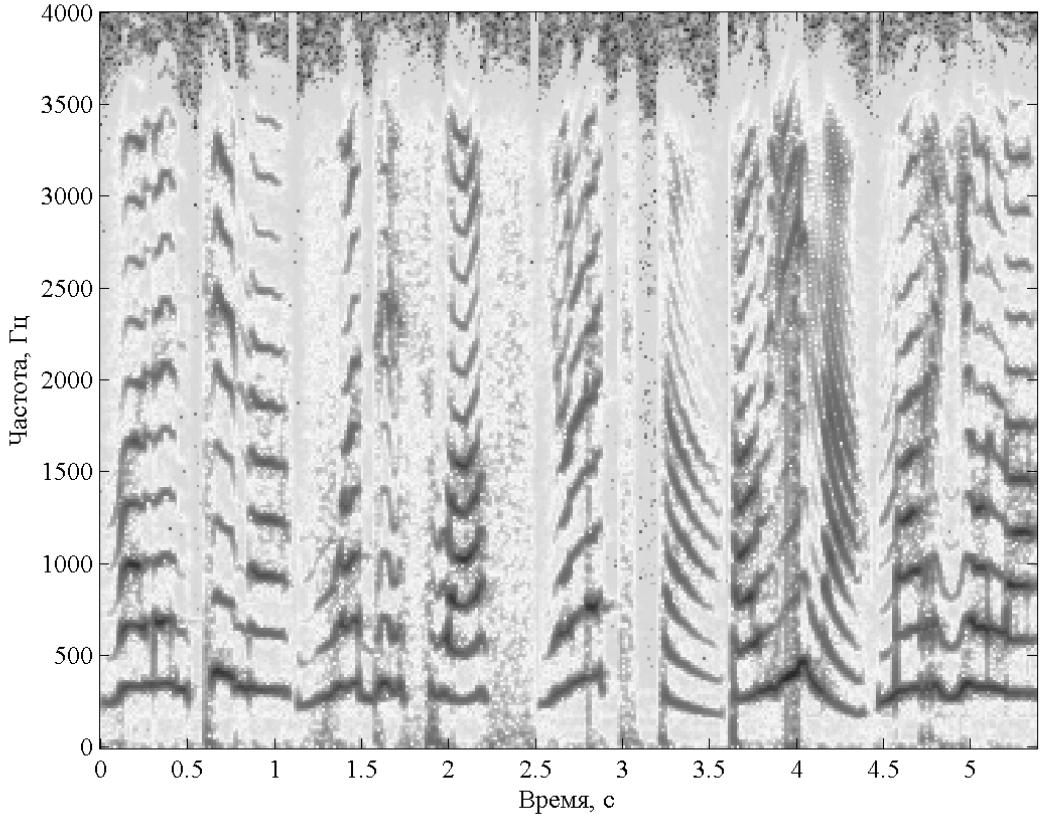


Рис. 2. Спектрограмма речевого фрагмента

Очевидно, что большая часть данного фрагмента имеет гармоническую структуру с фундаментальной частотой, изменяющейся в области примерно 300–400 Гц. Рассмотрим зависимость отношения "гармоники / шум" от предполагаемой частоты основного тона. Для этого выберем один из вокализованных сегментов данного фрагмента речи и проведем анализ по формулам (4)–(10) для типичных значений частоты основного тона речи от 50 до 500 Гц с шагом 1 Гц. Результат приведен на рис. 3.

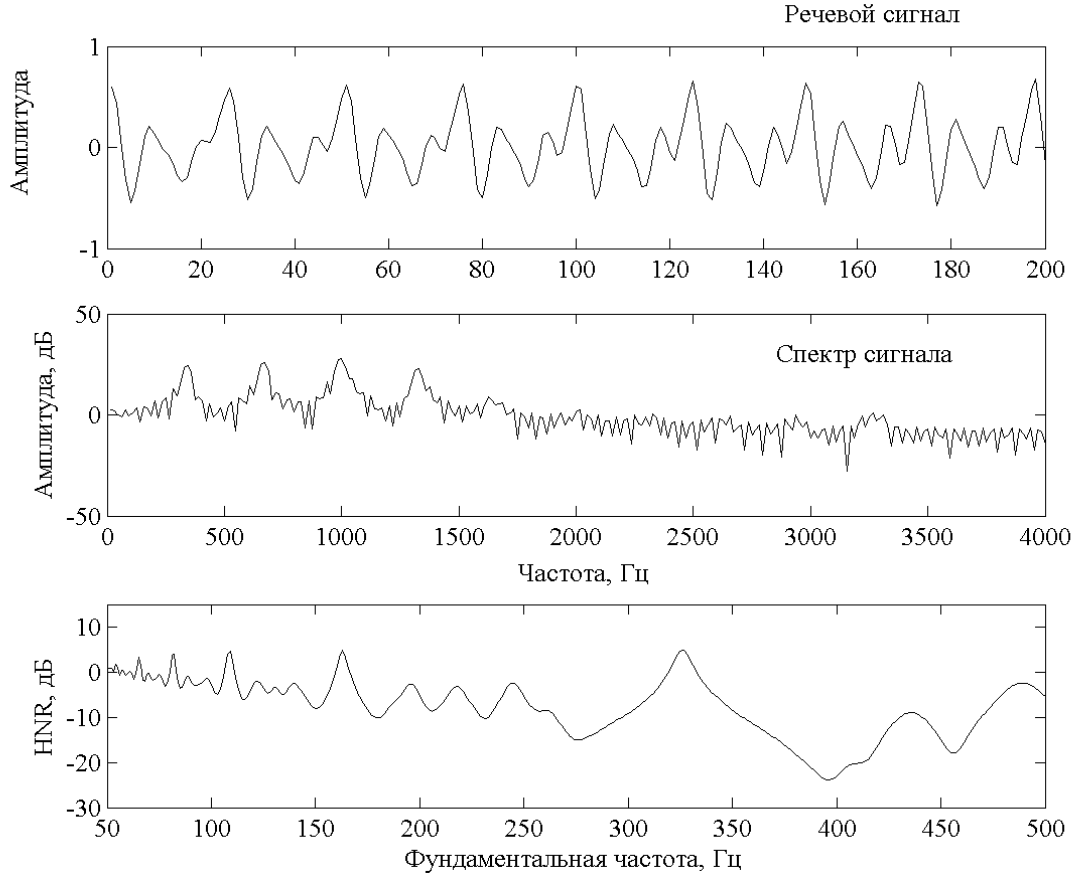


Рис. 3. Речевой сигнал, его спектр и зависимость отношения "гармоники / шум" от предполагаемого значения фундаментальной частоты

Из рис. 3 видно, что зависимость $HNR(F_0)$ имеет локальный максимум в точке, равной частоте основного тона данного сегмента речи и носит унимодальный характер в ее окрестности (область 280–390 Гц). Эксперименты показали, что характер данной зависимости является общим для вокализованных сегментов. Таким образом, параметры гармонической компоненты будут иметь оптимальные значения при

$$F_0^{opt} = \arg\max(HNR(F_0)),\ F_{0l} \le F_0 \le F_{0r}, \qquad (11)$$

где диапазон $[F_{0l}; F_{0r}]$ — некоторая окрестность фундаментальной частоты.

Таким образом, целесообразна следующая методика определения параметров гармонической компоненты: сначала проводится приблизительная оценка частоты основного тона и отслеживание ее контура, а затем в окрестности этой оценки проводится поиск максимума *HNR* методом анализа–через–синтез по формулам (4)–(10). Ниже приведен пример реализации данной методики:

В приведенном алгоритме используются следующие обозначения:

NRG — энергия сигнала на сегменте;

ThrNRG — заданное пороговое значение энергии вокализованного сегмента, определенное эмпирически;

NACF(k) — подпрограмма расчета нормализованной автокорреляционной функции по формуле

$$\psi(k) = \frac{\sum_{j=1}^{M} s_j s_{j+k}}{\sqrt{\sum_{j=1}^{M} s_j^2 \sum_{j=1}^{M} s_{j+k}^2}}, \qquad (12)$$

```
If NRG>ThrNRG then
begin
        For k=P_min to P_max do NACF(k);
        Candidates=Search_max_NACF(ThrAd);
        F0' =Track_DP(Candidates);
        if F0' >0 then
        begin
        Beta=Compute_Beta( F0' );
                for j=1 to n_iterations do
                        [ F0^opt , A, θ]=Golden_Section(F0l(j),F0r(j));
        end;
end;
```

где k — порядок автокорреляции; M — длина сегмента речи в отсчетах; s_j, s_{j+k} — отсчеты сигнала;

Search_max_NACF(ThrAd) – подпрограмма поиска максимумов нормализованной автокорреляционной функции (НАКФ), с которыми отождествляются кандидаты частоты основного тона Candidates. Рассматриваются только значения максимумов НАКФ, превышающие некоторый адаптивно изменяющийся порог ThrAd;

Track_DP(Candidates) — подпрограмма выбора траектории частоты основного тона методом динамического программирования [11]. Помимо кандидатов частоты основного тона в ней рассматривается и гипотеза о невокализованности речевого сегмента. В качестве параметров в данной подпрограмме используются значения НАКФ, с которыми отождествлены кандидаты частоты основного тона Candidates, а также расстояния между кандидатами частоты основного тона для смежных сегментов. Траектория частоты основного тона должна представлять собой для вокализованных звуков плавную линию. Результатом работы подпрограммы является приблизительная оценка фундаментальной частоты $F_0' > 0$ или вывод о невокализованности данного сегмента речи $F_0' = 0$;

Compute_Beta(F_0') — подпрограмма, осуществляющая расчет фактора изменения гармонических амплитуд β;

Golden_Section(F_{0l}(j),F_{0r}(j)) — подпрограмма, осуществляющая поиск максимума отношения "гармоники / шум" в окрестности значения F_0' методом "золотого сечения", выполняется n_iterations раз. На каждой итерации в подпрограмме осуществляется расчет векторов гармонических амплитуд A и фаз θ по формулам (5)–(7) с последующим синтезом гармонической компоненты по формуле (4) и определением нового значения HNR.

Определив оптимальное значение частоты основного тона F_0^{opt}, можно определить оптимальное значение ΔF_0, используя вышеприведенный алгоритм. Условием оптимальности будет следующее выражение:

$$\Delta F_0^{opt} = \arg\max(\mathrm{HNR}(F_0^{opt} + \Delta F_0)),\ \Delta F_{0l} \le \Delta F_0 \le \Delta F_{0r}, \qquad (13)$$

где диапазон $[\Delta F_{0l}; \Delta F_{0r}]$ представляет возможную область изменения частоты основного тона.

Итак, результатом работы представленного алгоритма является набор параметров, характеризующий гармоническую компоненту речевого сигнала: фундаментальная частота F_0, изменение фундаментальной частоты ΔF_0, вектор гармонических амплитуд A и вектор гармонических фаз θ, фактор β.

На рис. 4 показан результат работы приведенного алгоритма в составе схемы обработки речевого сигнала, представленной на рис. 1.

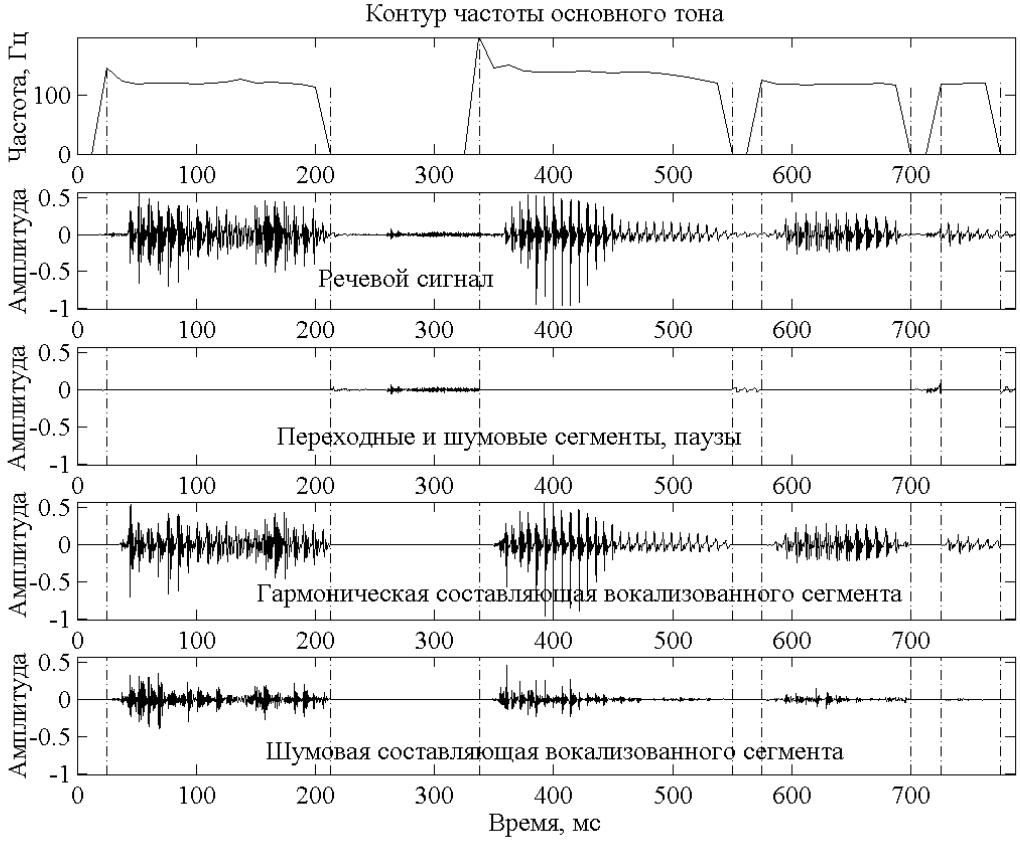
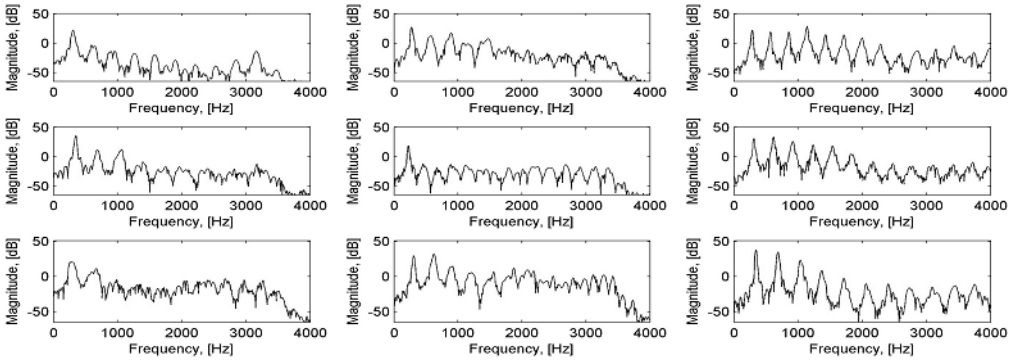


Рис. 4. Сепарация речевого сигнала.

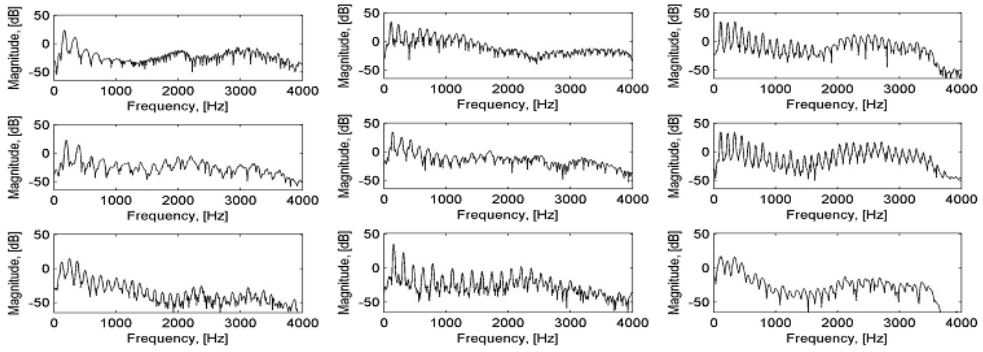
Квантование гармонических амплитуд

В системах анализа–синтеза и передачи речи задачи определения параметров и квантования имеют равную значимость. В контексте гармонической модели проблема квантования в большей степени связана с передачей вектора гармонических амплитуд. Если рассмотреть изменение спектра речевого сигнала во времени для разных дикторов (рис. 5), можно сделать вывод, что векторы гармонических амплитуд, даже определяющие голос одного и того же диктора, имеют различную размерность в разные моменты времени.

К сожалению, математический аппарат векторного квантования был разработан для квантования векторов фиксированной размерности и практически не используется с векторами переменной размерности, такими как векторы гармонических амплитуд. Для решения данной проблемы возможны различные подходы. Одним из вариантов является использование собственной кодовой книги для каждой размерности [12]. Естественно, такой подход является малоприемлемым для использования в системах реального времени из-за серьезных требований к объему памяти. Наиболее широко применяемые решения осуществляют различные преобразования над векторами переменной размерности с тем, чтобы привести их размерность к некоторому заданному фиксированному значению (с сохранением формы речевого спектра) с последующим применением техник векторного квантования. Примерами таких решений могут служить [9, 13–15]. Очевидным недостатком здесь является необходимость дополнительных преобразований и, следовательно, возможность внесения дополнительных искажений.



а



б

Рис. 5. Изменение спектра речи во времени: *а*) женский голос; *б*) мужской голос

Одно из возможных решений — квантование фиксированного количества гармонических амплитуд, например, в кодере на базе линейного предсказания со смешанным возбуждением (MELP — Mixed Excitation Linear Prediction) [16] векторное квантование используется для квантования первых 10 гармонических амплитуд, а амплитуды остальных гармоник считаются равными амплитуде последней (10-й) гармоники. Легко заметить, что 10 гармоник покрывают весь или почти весь речевой спектр для женских голосов с высокой частотой основного тона, в то время как для мужских голосов они могут покрыть только одну четвертую всего частотного диапазона (рис. 5,*а*, *б*), что означает существенную потерю качества для мужских голосов по сравнению с женскими.

Наконец, в [17] была разработана схема векторного квантования с переменной размерностью векторов (от англ. — Variable Dimension Vector Quantization — VDVQ). Тем не менее, поскольку в этом подходе не учитываются закономерности психоакустики, его трудно считать оптимальным.

Далее будет рассмотрен математический аппарат VDVQ и некоторые его модификации с точки зрения человеческого восприятия речи.

Векторное квантование с переменной размерностью векторов

В схеме VDVQ, предложенной в [17], кодовая книга квантователя содержит N_c кодовых векторов:

$$y_i,\ \ i = 0,....,N_c - 1 \tag{14}$$

при

$$y_i^T = \left[y_{i,0} \quad y_{i,1} \quad \cdots \quad y_{i,N_v-1} \right], \tag{15}$$

где N_v — размерность кодового вектора.

Пусть поиск вектора гармонических амплитуд x с размерностью $N(\omega_0)$ и нормализованной частотой основного тона ω_0 осуществляется путем полного перебора в кодовой книге, тогда требуется рассчитать следующие расстояния:

$$d_i(x, \hat{y}_i), \quad i = 0,..., N_c - 1, \tag{16}$$

где

$$\hat{y}_i^T = \left[\hat{y}_{i,1} \quad \hat{y}_{i,2} \quad \cdots \quad \hat{y}_{i,N(\omega_0)} \right], \tag{17}$$

$$\hat{y}_{i,j} = y_{i,k_j}, \; j = 1,..., N(\omega_0), \tag{18}$$

при

$$k_j = \left[\frac{N_V \omega_j}{\pi} \right], \; \omega_j = j\omega_0, \; j = 1,..., N(\omega_0), \tag{19}$$

где [.] означает округление к ближайшему целому.

Схема работает следующим образом: для каждого кодового вектора y_i путем расчета набора индексов k_j извлекается вектор $\hat{y}_i$, имеющий ту же размерность, что и x. Эти индексы рассчитываются в соответствии с периодом основного тона и указывают на элементы y_i, ближайшие к позиции j-й гармоники в кодовой книге. После расчета всех расстояний d_i для квантования x выбирается индекс кодового вектора с наименьшим расстоянием. В качестве расстояния (меры искажения) используется спектральное отклонение:

$$SD = \sqrt{\frac{1}{N(\omega_0)} \sum_{j=1}^{N(\omega_0)} \left(x_j - \hat{y}_j \right)^2}. \tag{20}$$

Улучшенная конфигурация схемы VDVQ, называемая IVDVQ, предложена в [18]. Улучшение основано на интерполяции элементов кодовых векторов y_i для получения действительных кодовых векторов $\hat{y}_i$. Индексы k_j в IVDVQ рассчитываются без операции округления:

$$k_j = \frac{N_V \omega_j}{\pi} \; \omega_j = j\omega_0, \; j = 1,..., N(\omega_0). \tag{21}$$

Элемент $\hat{y}_{i,j}$ получается путем линейной интерполяции между двумя элементами вектора y_i, определяемыми индексами $\lfloor k_j \rfloor$ и $\lceil k_j \rceil$:

$$\hat{y}_{i,j} = y_{i,\lfloor k_j \rfloor} + \{k_j\} \left(y_{i,\lceil k_j \rceil} - y_{i,\lfloor k_j \rfloor} \right), \tag{22}$$

где $\{k_j\}$ обозначает дробную часть выражения (21). Обучение кодовых книг по методам VDVQ и IVDVQ представляет собой вариацию на тему алгоритма "k-средних" [19] и подробно описано в [18]. Результат применения метода к квантованию гармонических амплитуд отражен на рис. 6, использовалась 10-разрядная кодовая книга.

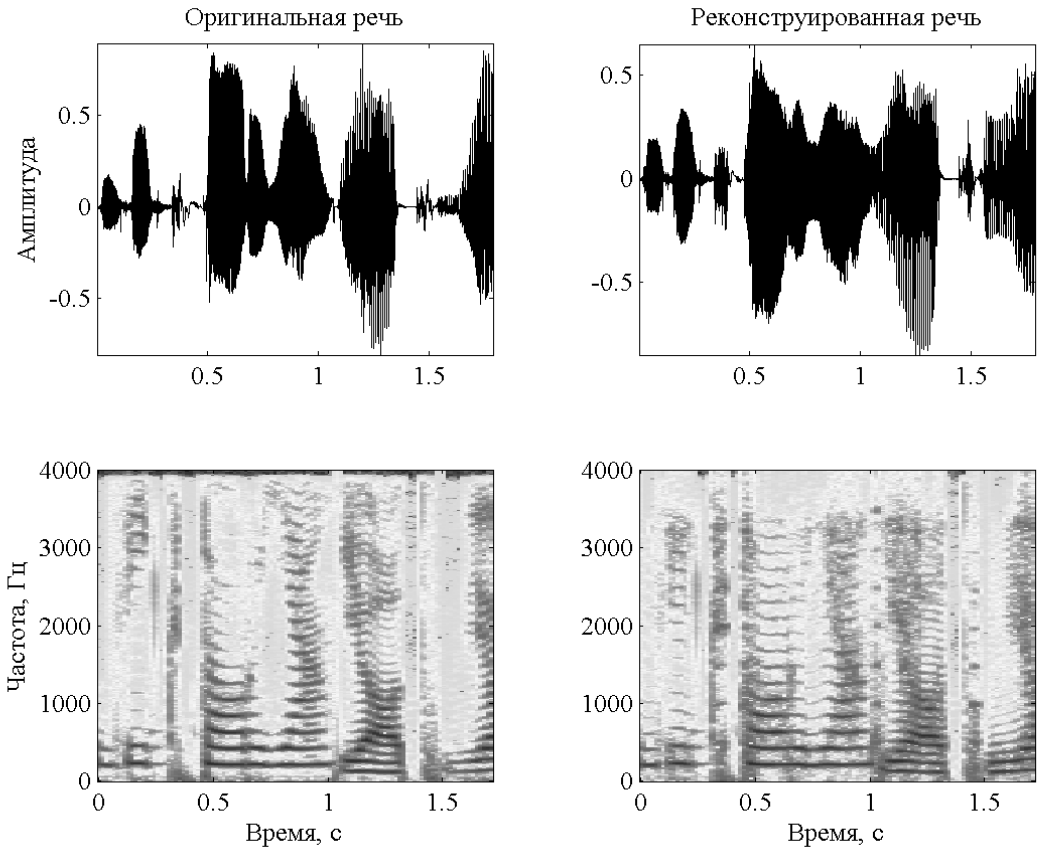


Рис. 6. Пример восстановления речи, кодированной с использованием метода VDVQ

VDVQ с применением линейной шкалы чувствительности слуховой системы человека

Метод квантования IVDVQ может рассматриваться как приемлемое решение для задачи квантования гармонических амплитуд. Если же посмотреть на эту задачу с точки зрения человеческого восприятия речи, становится очевидным, что IVDVQ не является оптимальным подходом. Данный вывод объясняется тем, что в процессе обучения кодовой книги в качестве критерия качества используется спектральное отклонение, не зависящее от частоты. В то же время человеческий слух имеет различную чувствительность к звукам разной частоты. Другим аспектом является нелинейная зависимость восприятия приращения громкости от приращения амплитуды. Таким образом, одно и то же численное значение разности между двумя гармониками может соответствовать совершенно разному перцептуальному искажению.

Для устранения таких несоответствий предлагается перцептуально обоснованный метод IVDVQ. Для того чтобы производить квантование гармонических амплитуд с учетом особенностей человеческого восприятия, величину амплитуд следует выражать не в децибелах, а в сонах и соответственно осуществлять обучение кодовой книги и поиск в ней (расчет расстояния между векторами – формула (20)). Шкала изменения громкости в сонах считается линейной для человеческого уха и определяется как [20]

$$A_s = 2^{\frac{A_p - 40}{10}}, \qquad (22)$$

где A_p и A_s — значения амплитуд, выраженные в фонах и сонах соответственно.

Единица измерения "фон" связана с единицей измерения "децибел" частотной характеристикой уха, значение в фонах определяется кривыми равной громкости [21], которые можно аппроксимировать следующим выражением, справедливым для речи средней громкости:

$$A_p = A_{dB} - ATH(f) + ATH_{1kHz}, \quad (23)$$

где A_p и A_{dB} — значения гармонических амплитуд в фонах и децибелах соответственно, ATH – функция, аппроксимирующая значение абсолютного порога слышимости [21]:

$$ATH(f) = 3,64 f^{-0.8} - 6,5 e^{-0.6(f-3.3)^2} + 10^{-3} f^4, \quad (24)$$

где f — частота в кГц.

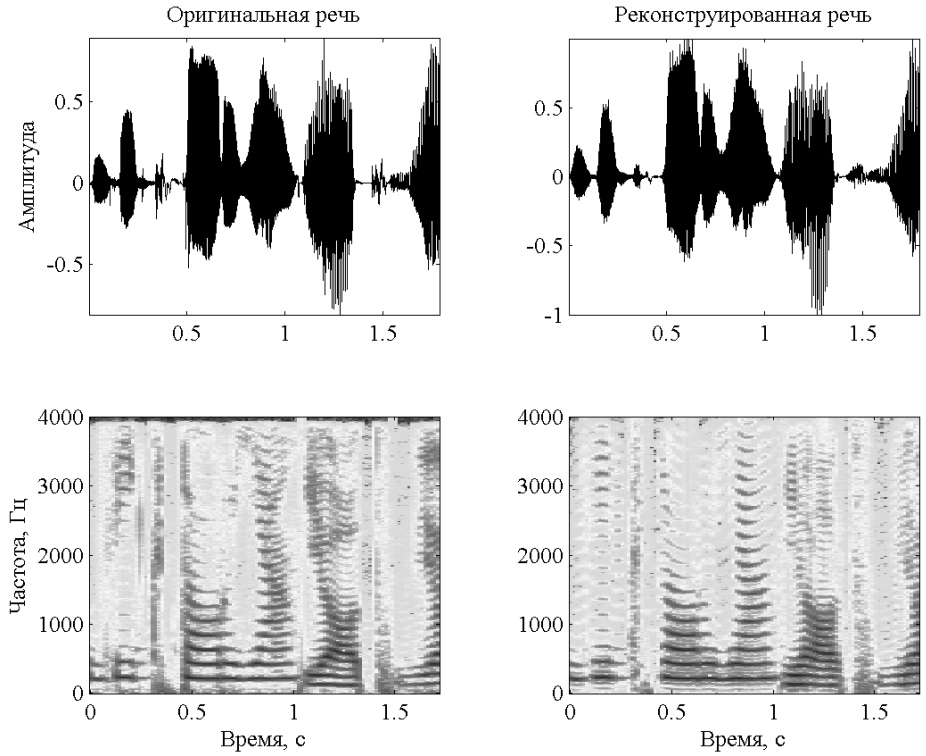


Рис. 7. Результат применения метода VDVQ, использующего линейную шкалу чувствительности слуховой системы человека

Таким образом, обучение кодовой книги и последующее квантование гармонических амплитуд основано на минимизации психоакустического искажения, что является преимуществом по сравнению с традиционным подходом VDVQ. Результат применения метода к квантованию гармонических амплитуд отражен на рис. 7, использовалась 10-разрядная кодовая книга.

VDVQ с психоакустически обоснованным ограничением длины вектора

Кодовые книги для VDVQ-метода обычно имеют большую длину кодовых слов (от 41 до 109 — в экспериментах [18]), что приводит к высоким требованиям к объему памяти для их хранения. В то же время можно видеть, что последние гармонические амплитуды спектра имеют незначительную величину, особенно в случае мужской речи (рис. 5б). Следовательно, имеет смысл ограничить размерность квантуемого вектора таким образом, чтобы не учитывать достаточно малые амплитуды. Схожая проблема существует в рамках модели речевого сигнала "гармоники плюс шум" [2, 22], где необходимо найти максимальную частоту вокализованности (ограничить спектр гармонической компоненты). Алгоритм, предложенный в [2], осуществляет проверку спектра на "гармоничность" в окрестности гармонических амплитуд, в случае, если спектр в области двух смежных проверяемых гармоник оказался "негармоническим", проверка прекращается. В качестве максимальной частоты вокализованности принимается последняя гармоника частоты основного тона,

предшествующая "негармонической" области спектра. Все же данный алгоритм является в большой степени эвристическим и использует при оценке некоторые заранее определенные опытным путем пороговые значения.

Модель анализа речевого сигнала, рассмотренная в данной статье, предполагает разделение речи на гармоническую и шумовую компоненту по всему спектру. Используя закономерности психоакустики, можно определить, в какой степени шумовая компонента влияет на восприятие человеком гармонической компоненты, т.е. определить гармоники, не влияющие на восприятие речи в целом.

Для решения данной проблемы использовалась психоакустическая модель Джонстона [23]. Данная модель позволяет рассчитать порог маскирования "шум маскирует тон" в частотной области с использованием следующей последовательности действий:

1. Сегмент шумовой компоненты, полученный с помощью выражения (9) взвешивается временным окном и подвергается ДПФ.

2. Спектр мощности шумовой компоненты суммируется в критических полосах, измеряемых в барках [21]:

$$B_i = \sum_{n=bl_i}^{bh_i} P(n), \tag{25}$$

где $P(n)$ — n-й частотный компонент спектра мощности; bl_i, bh_i — номера начального и конечного спектральных отсчетов, попадающих в i-ю критическую полосу.

Шкала барков получается с помощью следующего преобразования:

$$z(f) = 1 + 13\,\mathrm{arctg}(0,76f) + 3,5\,\mathrm{arctg}((f/7,5)^2), \tag{26}$$

где f — частота в Гц. Для *ДПФ размерности* 256 и частоты дискретизации F_s=8000 Гц параметры критических полос приведены в табл. 1.

3. Рассчитывается функция распространения для оценки эффектов маскирования в нескольких критических полосах [24]:

$$S_{i,j} = 10^{(15,81+7,5(k+0,474)-17,5\sqrt{1+(k+0,474)^2})/10}, \tag{27}$$

где $k=i-j$; i — номер барка маскируемого сигнала; j — номер барка маскирующего сигнала.

4. Вычисляется распространение спектральной энергии барка в каждой критической полосе как свертка B_i с функцией распространения $S_{i,j}$:

5.

$$\begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ \dots \\ C_{18} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & \dots & S_{1,18} \\ S_{2,1} & S_{2,2} & S_{2,3} & \dots & S_{2,18} \\ S_{3,1} & S_{3,2} & S_{3,3} & \dots & S_{3,18} \\ \dots & \dots & \dots & \dots & \dots \\ S_{18,1} & S_{18,2} & S_{18,3} & \dots & S_{18,18} \end{bmatrix} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \dots \\ B_{18} \end{bmatrix}, \tag{28}$$

6. Рассчитываются коэффициенты тональности для каждой критической полосы:

$$\alpha_i = \min\left(\frac{SFM_{dB}(i)}{SFM_{dB\max}}, 1\right), \tag{29}$$

где $SFM_{dB}(i)$ — мера спектральной пологости в i-й критической полосе:

$$SFM_{dB} = 10\left[\log_{10}(GM) - \log_{10}(AM)\right], \tag{30}$$

где AM и GM — среднее арифметическое и среднее геометрическое значение спектра мощности в i-й критической полосе; SFM_{dBmax} — максимальное значение меры спектральной пологости равное –60 дБ.

Таблица 1. **Параметры критических полос приведены для *ДПФ размерности* 256 и частоте дискретизации F_s=8000 Гц**

Номер критической полосы	Номера элементов ДПФ	Количество элементов ДПФ	Частоты, Гц
1	1–3	3	0–94
2	4–6	3	94–187
3	7–10	4	187–312
4	11–13	3	312–406
5	14–16	3	406–500
6	17–20	4	500–625
7	21–25	5	625–781
8	26–29	4	781–906
9	30–35	6	906–1094
10	36–41	6	1094–1281
11	42–47	6	1281–1469
12	48–55	8	1469–1719
13	56–64	9	1719–2000
14	65–74	10	2000–2312
15	75–86	12	2312–2687
16	87–100	14	2687–3125
17	101–118	18	3125–3687
18	119–128	9	3687–4000

7. Определяются смещения порогов маскирования:

$$O_i = 5,5(1-\alpha_i). \quad (31)$$

8. Производится расчет порогов маскирования в критических полосах и их ренормализация:

$$T_i = 10^{\log_{10}(C_i)-O_i/10}. \quad (32)$$

Для ренормализации требуется определить ошибку распространения спектральной энергии барка, для этого предполагается, что на слуховую систему воздействует гипотетический раздражитель, спектральная энергия которого в критической полосе равна единице:

$$\begin{bmatrix} C_{E1} \\ C_{E2} \\ C_{E3} \\ \dots \\ C_{E18} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & \dots & S_{1,18} \\ S_{2,1} & S_{2,2} & S_{2,3} & \dots & S_{2,18} \\ S_{3,1} & S_{3,2} & S_{3,3} & \dots & S_{3,18} \\ \dots & \dots & \dots & \dots & \dots \\ S_{18,1} & S_{18,2} & S_{18,3} & \dots & S_{18,18} \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}. \quad (33)$$

Ренормализованные пороги маскирования определяются как

$$T_i^{'} = T_i - 10\log_{10}(C_{Ei}). \quad (34)$$

9. Окончательные значения порогов маскирования определяются как

$$T_i^{f} = \max\left(T_i^{'}, ATH(f)\right), \quad (35)$$

где *ATH*(*f*) рассчитывается с помощью выражения (24) для частот, равных значениям гармоник частоты основного тона.

Максимальной частотой вокализованности считается последняя гармоника частоты основного тона, превышающая порог маскирования.

На рис. 8 показан результат расчета порога маскирования и определения максимальной частоты вокализованности для вектора гармонических амплитуд. Очевидно, что

вычислительная сложность поиска в кодовой книге в данном случае будет снижена более чем в 2 раза.

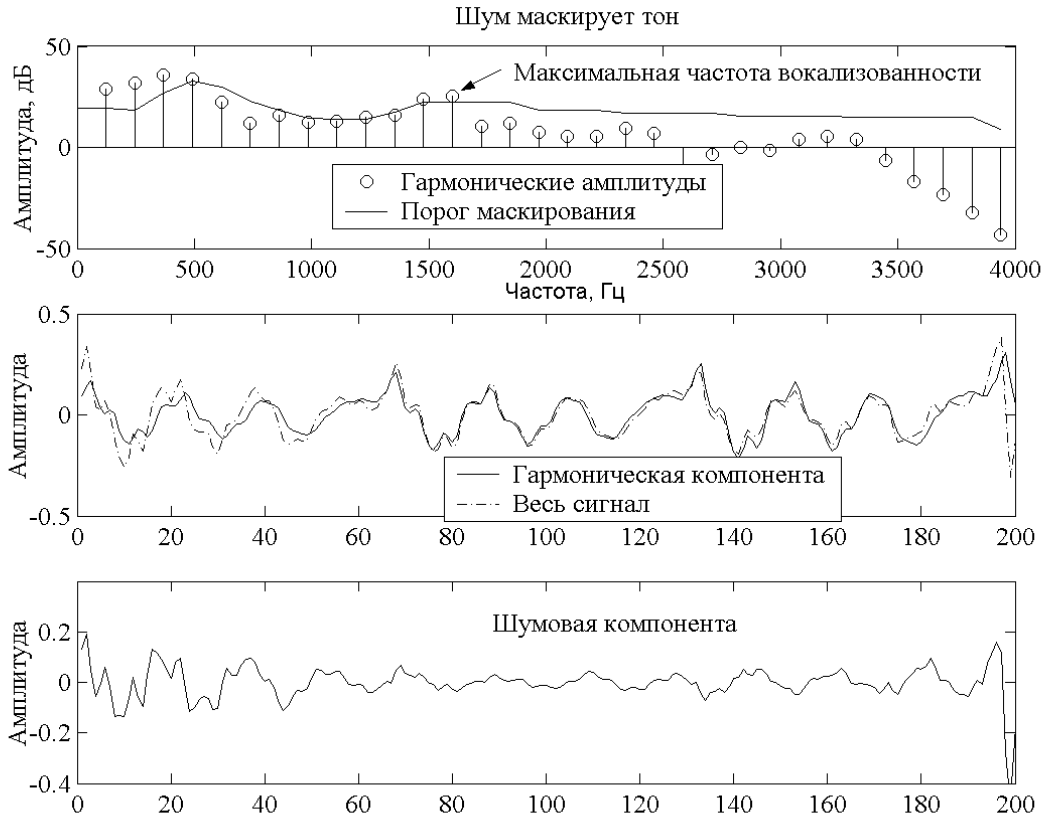


Рис. 8. Маскирование гармонических амплитуд.

Таким образом, удается ограничить размерность вектора гармонических амплитуд на основании закономерностей психоакустики и тем самым снизить вычислительную сложность процесса квантования гармонических амплитуд. Результат применения метода отражен на рис. 9, использовалась 10-разрядная кодовая книга.

Сравнительные результаты методов квантования векторов переменной размерности

Поскольку предлагаемые методы квантования основаны на использовании особенностей слуха человека, классические параметры, по которым можно их сравнить (отношение "сигнал/шум", спектральное отклонение и т.д.), не смогут обеспечить корректную оценку качества. В то же время оценка качества по шкале MOS (Mean Opinion Score) требует наличия специально оборудованного помещения и определенного количества подготовленных слушателей. Таким образом, целесообразным будет произвести оценку качества реконструированной речи с помощью такого параметра, при расчете которого использовалась бы модель слуха человека. Таким параметром является модифицированная величина искажений спектра барков (MBSD — Modified Bark Spectral Distortion) [25], искажения в данном случае определяются как средняя разность субъективных оценок громкости.

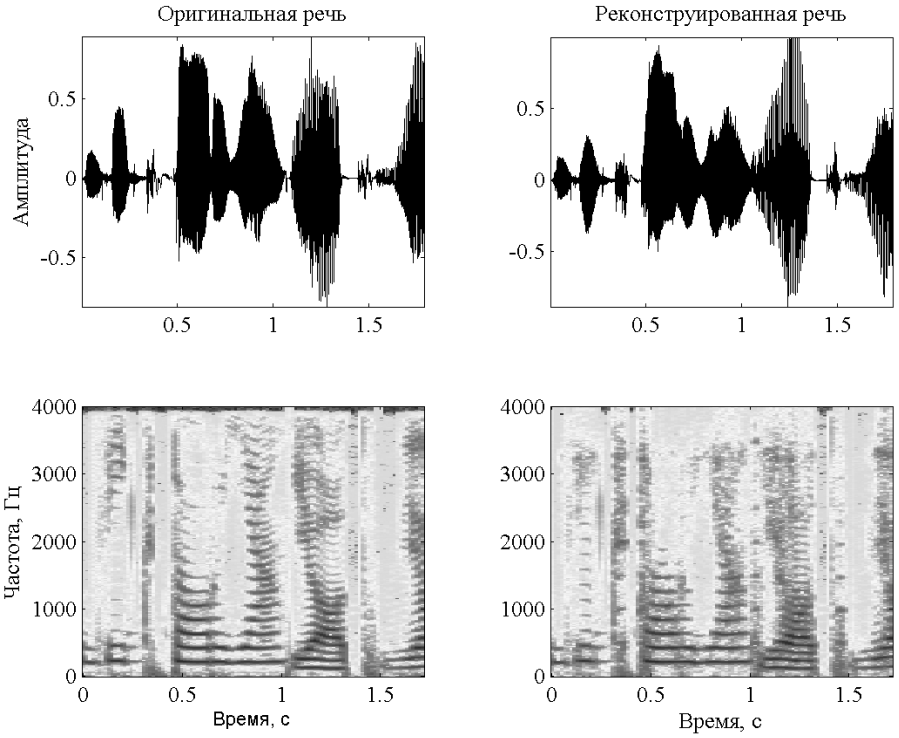


Рис. 9. Результат применения метода VDVQ с психоакустически мотивированным ограничением длины вектора.

Для оценки качества использовались десятиразрядные кодовые книги, полученные с использованием подходов, описанных выше. Результаты тестирования качества реконструированной речи для различных вариантов квантования гармонических амплитуд приведены в табл. 2.

Таблица 2. **Качество реконструированной речи при использовании различных подходов для квантования векторов гармонических амплитуд**

	VDVQ	VDVQ+преобразование "децибелы-соны"	VDVQ+психоакустически обоснованное ограничение длины вектора
MBSD	5,5973	2,3769	5,3348

Таким образом, психоакустически модифицированные варианты квантования векторов гармонических амплитуд показали по результатам измерений лучшее качество с точки зрения восстановления речи. Самый лучший результат был обеспечен при использовании метода VDVQ с применением линейной шкалы чувствительности слуховой системы человека.

Заключение

В данной статье была рассмотрена гармоническая модель речевого сигнала с точки зрения определения ее параметров и последующего их квантования.

При определении параметров гармонической модели использовался математический аппарат дискретного преобразования Фурье, согласованного с изменением контура частоты основного тона. Данный подход обеспечивает высокую точность результатов при условии корректного определения фундаментальной частоты. Применение техники динамического программирования в совокупности с последующим одновременным уточнением частоты основного тона и определением параметров гармонической модели в цикле с обратной связью методом анализа–через–синтез позволяет добиться качественного разделения речевого сигнала на периодическую (непосредственно гармоническую) и апериодическую (шумовую) компоненты. Сепарация речи на эти две разные по своей природе составляющие, а

следовательно, описание их разными наборами параметров, приводит к уменьшению пространства состояний для процедуры квантования, что, в свою очередь, упрощает подготовку кодовых книг и улучшает их качество.

Метод квантования векторов переменной размерности является весьма удобным для использования с такими параметрами гармонической модели речи как амплитуды, поскольку отпадает надобность в дополнительных преобразованиях. Предложенные методы, в основе которых лежат преобразования, использующие закономерности психоакустики позволяют повысить качество реконструированной речи и снизить вычислительную сложность алгоритмов квантования. Возможно, лучшие результаты даст объединение подходов, связанных с преобразованием шкал "децибелы-соны" и психоакустически обоснованным ограничением длины вектора гармонических амплитуд.

HARMONIC MODEL OF THE SPEECH SIGNAL: PARAMETERS ESTIMATION AND QUANTIZATION

A.N. PAVLOVETS, P. ZUBRYCKI, A.A. PETROVSKY

Abstract

The method of estimation of parameters of speech signal harmonic model is considered in this paper. The main feature of this method is application of the closed-loop analysis-by-synthesis algorithm for joint pitch refinement and harmonic amplitudes computing. Further quantization of harmonic amplitudes vector is improved by perceptually based methods.

Литература

1. *Almeida L., Tribolet J.* // IEEE Trans. on Acoust., Speech, Sig. Proc. 1983. Vol. ASSP-31, № 3. P. 664–678.
2. *Stylianou Y.* // IEEE Trans. on Speech and Audio Proc. 2001. Vol. 9, № 1. P. 21–29.
3. *Shlomot E., Cuperman V., Gersho A.* // IEEE Trans. Speech and Audio Proc. 2001. Vol. 9, № 6. P. 632–646.
4. *Griffin D., Lim J.* // IEEE Trans. on Acoust., Speech, Sig. Proc. 1988. Vol. 36, №8. P. 1223–1235.
5. *Петровский А.А., Серков В.В.* // Цифровая обработка сигналов. 2002. № 2. С. 2–12.
6. *Petrovsky A., Zubricki P., Savicki A.* // Proc. Europ. Conf. on Circuit Theory and Design. 2003. Vol. 3. P. 169–172.
7. *Sercov V., Petrovsky A.* // Proc. EUROSPEECH'99. 1999. P. 1479–1482.
8. *Gersho A., Gray R.M.* Vector Quantization and Signal Compression. Kluwer Academic, Norwell, USA. 1992.
9. *Павловец А.Н., Петровский А.А.* // Цифровая обработка сигналов. 2005. № 3. С. 13–21.
10. *Jensen J., Jensen S., Hansen E.* // Proc. IEEE ICASSP'2000. 2000. P. 1439–1442.
11. *Talkin D.* // Speech Coding and Synthesis. Editors: W.B. Kleijn and K.K. Palival. Elsevier. Amsterdam, Netherlands. 1995.
12. *Adoul J.-P., Delprat M.* // Proc. Allerton Conf. on Circuits, Syst., Comput. 1986. P. 1004–1011.
13. *McAulay R.J., Quatiery T.F.* // Speech Coding and Synthesis. Editors: W.B. Kleijn, K.K. Palival. Elsevier, Amsterdam, Netherlands. 1995.
14. *Nishiguchi M., Inoue A., Maeda Y., Matsumoto J.* // Proc. IEEE Speech Coding Workshop. 1999. P. 84–86.
15. *Li C., Lupini P., Shlomot E., Cuperman V.* // IEEE Trans. on Speech and Audio Proc. 2001. Vol. 9, № 6. P. 622–631.
16. *Supplee L., Cohn R., Collura J., McCree A.* // Proc. IEEE ICASSP'97. 1997. Vol. 2, P. 1591–1594.
17. *Das A., Rao A., Gersho A.* // IEEE Sig. Proc. Letters. 1996. Vol. 3. № 7. P. 200–202.
18. *Chu W.* // Proc. 3rd IEEE Int. Symp. on Image and Sig. Proc. and Analysis. 2003. Vol. 1. P. 537–542.
19. *MacQueen, J.* // Proc. 5th Berkeley Symp. on Math. Stat. and Prob. 1. 1967. P. 281–296.
20. *Bladon R.* // J. of the Acoust. Soc. of America. 1981. Vol. 69. P. 1414–1422.
21. *Zwicker E., Fastl H.* / Psychoacoustics: facts and models. Berlin: Springer-Verlag, 1990.
22. *Bao C., Lukasiak J., Ritz C.* // Proc. Interspeech'2005. 2005. P. 2709–2712.
23. *Johnston, J.* // Proc. IEEE ICASSP'88. 1988. A1.9, P. 2524–2527.
24. *Schroeder M. R., Atal B. S., Hall J. L.* // J. of the Acoust. Soc. of America. 1979. Vol. 66. P. 1647–1652.
25. *Yang W., Benbouchta M., Yantorno R.* // Proc. IEEE ICASSP'98. 1998. P. 541–544.