Использование закономерностей психоакустики в процедуре квантования параметров гармонической модели речевого сигнала

А.Н. Павловец,
аспирант
А.А. Петровский,
доктор технических наук, профессор

В данной работе рассматривается метод векторного квантования с переменной размерностью векторов для параметров гармонической модели речевого сигнала. Особенностью метода является использование закономерностей психоакустики в процедуре квантования амплитуд, что позволяет повысить качество реконструированного речевого сигнала и снизить вычислительную сложность алгоритмов квантования.

Abctract
The method of variable dimension vector quantization of harmonic model parameters is considered in this paper. The essence of the method is the incorporation of psychoacoustic principles into quantization procedure.

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 80-х годов [1].

Одним из фундаментальных вопросов в приложениях кодирования на базе гармонических моделей является квантование амплитуд гармоник, 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 к кодированию последовательности амплитуд гармоник. Скалярное квантование, например, квантует каждый элемент индивидуально; тем не менее, векторное квантование [2] является более предпочтительным подходом для современных алгоритмов низкоскоростных кодеров речи, что обусловлено лучшим соотношением качество/скорость передачи. Традиционные векторные квантователи строятся с учётом фиксированной длины векторов. В последних работах удалось добиться достаточно высокого качества квантования гармонических амплитуд благодаря применению схемы расщеплённого векторного квантования линейных спектральных пар, при этом прозрачное кодирование достигалось при скорости 23 бит/вектор [3]. Тем не менее, построение векторного квантователя с переменной длиной кодируемого вектора амплитуд гармоник выглядит более естественным решением ввиду того, что при этом не требуется дополнительных преобразований над входным вектором.

Использование особенностей слуховой системы человека при низкоскоростном кодировании речи было рассмотрено в работе [4], где закономерности психоакустики учитывались при построении огибающих спектров и при расчёте весовых коэффициентов для взвешивания ошибки квантования параметров в контексте MBE-вокодера. Таким образом, целью данной работы является разработка метода квантования векторов амплитуд гармоник с учётом особенностей восприятия речи человеком.

Квантование гармонических амплитуд

В контексте гармонической модели проблема квантования в большей степени связана с передачей вектора амплитуд гармоник. Если рассмотреть изменение спектра речевого сигнала во времени для разных дикторов (***рис. 1 а и 1 б***), можно сделать вывод, что векторы амплитуд гармоник, 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 длины и практически не используется с векторами переменной длины, такими, как векторы амплитуд гармоник. К решению данной проблемы возможны различные подходы. Одним из вариантов является использование собственной кодовой книги для каждой размерности [5]. Естественно, такой подход является малоприемлемым для использования в системах реального времени из-за серьёзных требований к объёму памяти. В наиболее широко применяемых решениях осуществляются 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 [3, 6–8]. Очевидным недостатком здесь является необходимость дополнительных преобразований и, следовательно, возможность внесения дополнительных искажений.

Одно из возможных решений — квантование фиксированного количества гармонических амплитуд: например, в кодере на базе линейного предсказания со смешанным возбуждением

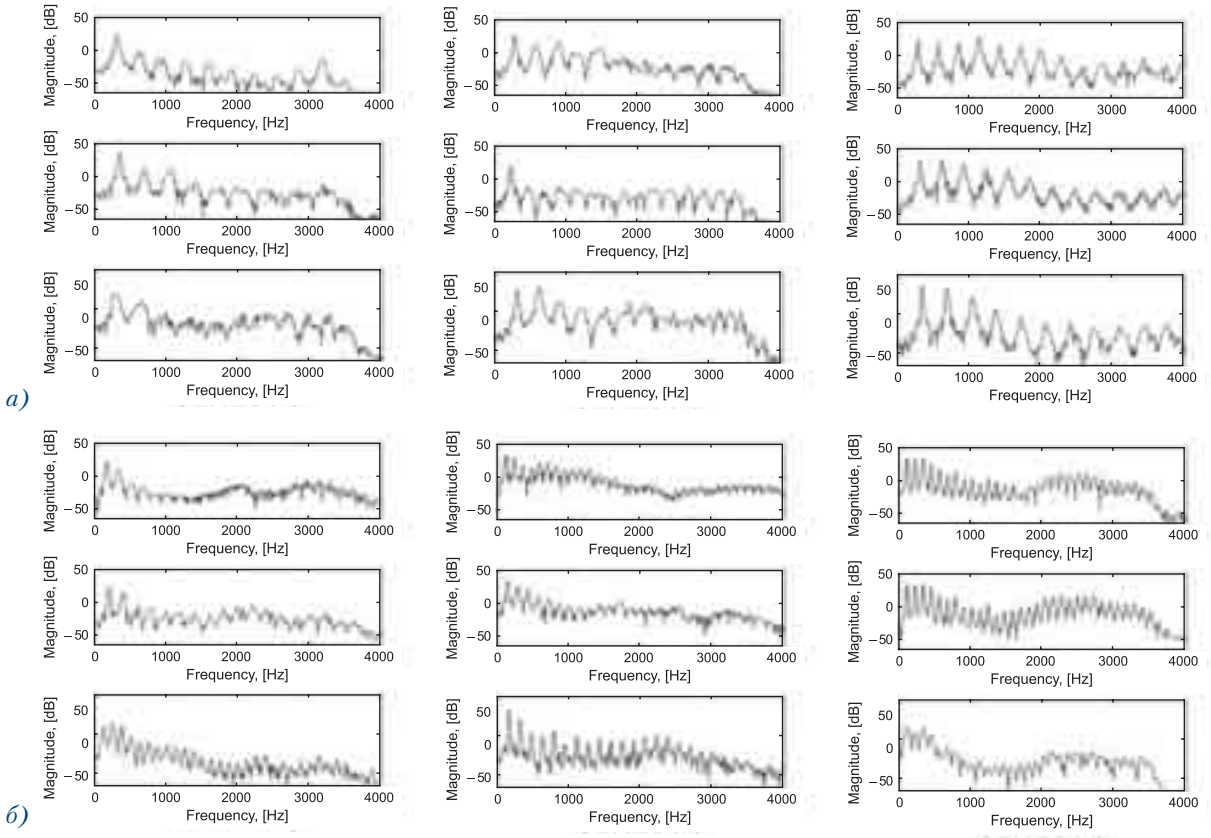

Рис. 1. Изменение спектра речи во времени: а) женский голос; б) мужской голос

(MELP — Mixed Excitation Linear Prediction) [9] векторное квантование используется для квантования первых 10-ти 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. 1а и 1б***), что означает существенную потерю качества для мужских голосов по сравнению с женскими.

Наконец, в [10] 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 его модификация с точки зрения восприятия речи человеком.

Векторное квантование с переменной размерностью векторов

В схеме VDVQ, предложенной в [10], кодовая книга квантователя содержит N_c кодовых векторов:

$$y_i,\ i = 0,...,N_c - 1$$

при

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

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 путём полного перебора в кодовой книге, тогда требуется рассчитать следующие расстояния:

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

где

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

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

при

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

где [] означает округление к ближайшему целому.

Схема работает следующим образом: для каждого кодового вектора y_i путём расчёта 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 расчёта всех расстояний d_i для квантования x выбирается индекс кодового вектора с наименьшим расстоянием. В качестве расстояния (меры искажения) используется спектральное отклонение:

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

Модифицированная конфигурация схемы VDVQ, называемая IVDVQ, предложена в [11]. Данное изменение заключается в интерполяции элементов кодовых векторов y_i для получения действительных кодовых векторов $\hat{y}_i$. Индексы k_j в IVDVQ рассчитываются без операции округления:

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

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

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

где $\{k_j\}$ обозначает дробную часть выражения (1). Обучение кодовых книг по методам VDVQ и IVDVQ представляет собой вариацию на тему алгоритма «k-средних» [12] и подробно описано в [11]. Результат применения метода к квантованию гармонических амплитуд отражён на ***рис. 2***; использовалась 10-разрядная кодовая книга.

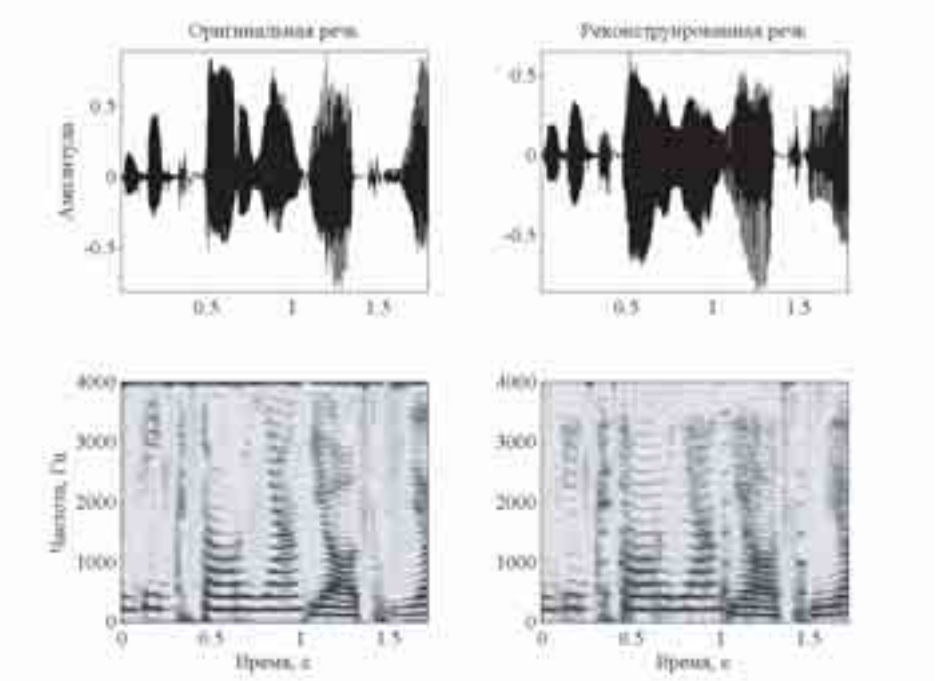

Рис. 2. Пример восстановления речи, кодированной с использованием метода VDVQ

VDVQ с психоакустически обоснованным ограничением длины вектора

Кодовые книги для VDVQ-метода обычно имеют большую длину кодовых слов (от 41 до 109 — в экспериментах [11]), что приводит к высоким требованиям к объёму памяти для их хранения. В то же время можно видеть, что последние амплитуды гармоник спектра имеют незначительную величину, особенно в случае мужской речи (рис. 1б). 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» [13, 14], где необходимо найти максимальную частоту вокализованности (ограничить спектр гармонической компоненты). Алгоритм, предложенный в [13], осуществляет проверку спектра на гармоничность в окрестности амплитуд гармоник, и 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. Всё же данный алгоритм является в большой степени эвристическим и использует при оценке некоторые заранее определённые опытным путём пороговые значения.

Модель анализа речевого сигнала, рассмотренная в [15], 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 [16]. Данная модель позволяет рассчитать порог маскирования «шум маскирует тон» в частотной области с использованием следующей последовательности действий:

1) Сегмент шумовой компоненты взвешивается временным окном и подвергается ДПФ;

2) Спектр мощности шумовой компоненты суммируется в критических частотных полосах, измеряемых в барках [17]:

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

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

Шкала барков получается с помощью следующего преобразования [17]:

$$z(f) = 1 + 13arctg(0{,}76f) + 3.5arctg((f/7{,}5)^2),$$

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

3) Рассчитывается функция распространения возбуждения по базилярной мембране для оценки эффектов маскирования в нескольких критических частотных полосах [18]:

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

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

4) Вычисляется распространение спектральной энергии барка в каждой критической частотной полосе как свёртка спектра мощности B_i с функцией распространения возбуждения $S_{i,j}$:

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

5) Рассчитываются коэффициенты тональности для каждой критической частотной полосы:

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

где $SFMdB(i)$ — мера спектральной пологости в i-ой критической частотной полосе:

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

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

Таблица 1

Параметры критических частотных полос приведены для ДПФ размерности 256 и частоты дискретизации Fs=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

6) Определяются смещения порогов маскирования:

$$O_i = 5{,}5(1-\alpha_i),\ i=1\ldots18.$$

7) Производится расчёт порогов маскирования в критических полосах и их ренормализация:

$$T_i = 10^{\log_{10}(C_i)-O_i/10},\ i=1\ldots18.$$

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

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

Ренормализованные пороги маскирования определяются так:

$$T_i^{'} = T_i - 10\log_{10}(C_{Ei}),\ i=1\dots18.$$

8) Окончательные значения порогов маскирования определяются так:

$$T_i^f = \max\left(T_i^{'}, ATH(f)\right),\ i=1\dots18,$$

где $ATH(f)$ — функция, аппроксимирующая значение абсолютного порога слышимости [17] и рассчитываемая с помощью следующего выражения для частот, равных значениям гармоник частоты основного тона:

$$ATH(f) = 3.64 f^{-0.8} - 6.5e^{-0.6(f-3.3)^2} + 10^{-3} f^4,$$

где f — частота в кГц.

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

На ***рис. 3*** показан результат расчёта порога маскирования и определения максимальной частоты вокализованности для вектора амплитуд гармоник. Очевидно, что вычислительная сложность поиска в кодовой книге в данном случае будет снижена более чем в 2 раза.

Таким образом удаётся ограничить размерность вектора амплитуд гармоник на основании закономерностей психоакустики и тем самым снизить вычислительную сложность процесса его квантования. Результат применения метода отражён на ***рис. 4***; использовалась 10-разрядная кодовая книга.

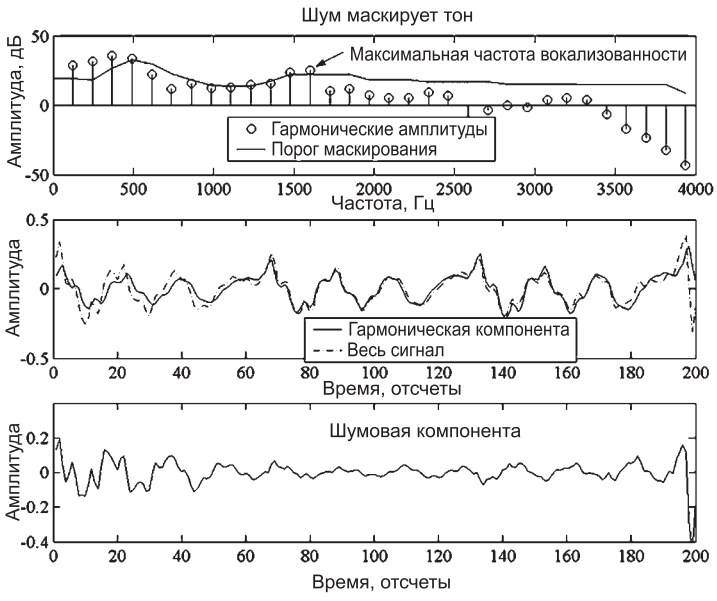

Рис. 3. Маскирование амплитуд гармоник шумовой компонентой

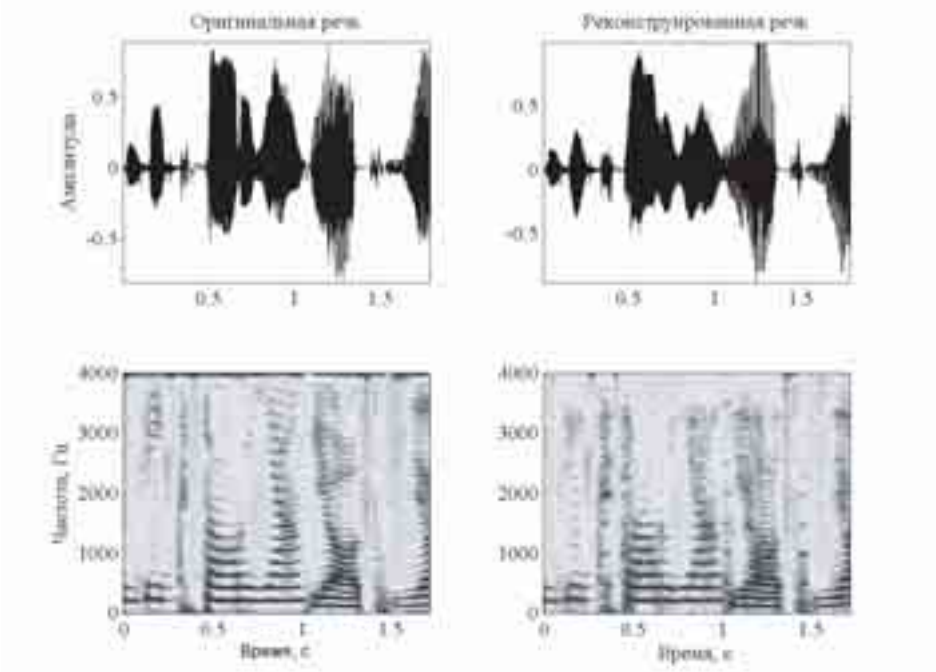

Рис. 4. Результат применения метода VDVQ с психоакустически мотивированным ограничением длины вектора

Сравнение качества методов квантования векторов амплитуд гармоник переменной длины

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 определённого количества подготовленных слушателей. Таким образом, целесообразным будет произвести оценку качества реконструированной речи с помощью такого параметра, при расчёте которого использовалась бы модель слуха человека. Таким параметром является модифицированная величина искажений спектра барков (MBSD — Modified Bark Spectral Distortion) [19], искажения в данном случае определяются как средняя разность субъективных оценок громкости.

Для сравнительной оценки качества квантователи, построенные на базе предложенных методов, были использованы в составе вокодера, основанного на декомпозиции речевого сигнала на периодическую и апериодическую компоненты [15]. В ходе эксперимента квантованию подвергались только амплитуды гармоник (использовались десятиразрядные кодовые книги), прочие параметры не квантовались. Результаты тестирования качества реконструированной речи для различных вариантов квантования приведены в ***таблице 2***.

Таблица 2

Качество реконструированной речи при использовании различных подходов для квантования векторов амплитуд гармоник

	VDVQ	VDVQ+психоакустически обоснованное ограничение длины вектора
MBSD	5,5973	5,3348

Таким образом, психоакустически модифицированный вариант квантования векторов амплитуд гармоник показал по результатам измерений лучшее качество с точки зрения восстановления речи.

Заключение

В данной статье были рассмотрены методы квантования векторов амплитуд гармоник речевого сигнала. Метод квантования векторов переменной длины 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.

Литература

1. *Almeida L., Tribolet J.* «Nonstationary spectral modeling of voiced speech», IEEE Trans. Acoustics, Speech, Signal Processing., vol.ASSP-31, №3, pp. 664–678, June 1983.
2. *Gersho A., Gray R.M.* Vector Quantization and Signal Compression. Kluwer Academic, Norwell, USA, 1992.
3. *Павловец А., Петровский А.* «Квантование огибающей спектра в вокодере, основанное на декомпозиции речевого сигнала на периодическую и апериодическую составляющие», Цифровая обработка сигналов, №3, Москва, 2005 г., с. 13–21.
4. *Серков В., Петровский А.* «Использование закономерностей психоакустики при низкоскоростном кодировании речи», Доклады 3-й междунар. конф. «Цифровая обработка сигналов и её применения», DSPA'2000, Москва, 2000, с. 241–244.
5. *Adoul J.-P., Delprat M.* «Design algorithm for variable-length vector quantizers» in Proc. Allerton Conf. on Circuits, Syst., Comput, 1986, pp. 1004–1011.
6. *McAulay R.J., Quatiery T.F.* «Sinusoidal Coding» in «Speech Coding and Synthesis» (W.Klein and K. Palival, eds.), Amsterdam: Elsevier Science Publishers, 1995, pp. 121–176.
7. *Nishiguchi M., Inoue A., Maeda Y., and Matsumoto J.* «Parametric speech coding-HVXC at 2.0–4.0 kbps» in Proc. IEEE Speech Coding Workshop, pp. 84–86, Porvoo, Finland, June 1999.
8. *Li C., Lupini P., Shlomot E., and Cuperman V.* «Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization», IEEE Trans. Speech, and Audio Processing, vol. 9, no. 6, pp. 622–631, 2001.
9. *Supplee L., Cohn R., Collura J., McCree A.* «MELP: the new federal standard at 2400 bps», in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'97, vol.2, pp. 1591–1594, Munich, Germany, April 1997.
10. *Das A., Rao A., Gersho A.* «Variable-dimension vector quantization», IEEE Signal Processing Letters, vol.3, no.7, pp. 200–202, 1996.

11. *Chu W.C.* «A novel approach to variable dimension vector quantization of harmonic magnitudes», in Proc. 3rd IEEE International Symposium on Image and Signal Processing and Analysis, vol.1, pp.537–542, Rome, Italy, September 2003.
12. *MacQueen, J.B.* «Some Methods for Classification and Analysis of Multivariate Observations», in Proc. Fifth Berkeley Symp. Math. Statistics and Probability, vol.1, pp. 281–296, 1967.
13. *Stylianou Y.* «Applying the harmonic plus noise model in concatenative speech synthesis», IEEE Transactions on Speech and Audio Processing, vol.9, №1, pp. 21–29, Jan. 2001.
14. *Bao C., Lukasiak J., Ritz C.* «A novel voicing cut-off determination for low bit-rate harmonic speech coding», in INTERSPEECH-2005, pp. 2709–2712.
15. *Petrovsky A., Zubricki P., Sawicki A.* «Tonal and noise components separation based on a pitch synchronous DFT analyzer as a speech coding method», in Proc. European Conf. Circuit Theory and Design, Cracow, Poland, Sep. 2003, vol.3, pp.169–172.
16. *Johnston J.* «Estimation of perceptual entropy using noise masking criteria», in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'88, vol.5, pp. 2524–2527, New York, NY, USA, April 1988.
17. *Zwicker E., Fastl H.* «Psychoacoustics: facts and models», Springer-Verlag, Berlin, 1990.
18. *Schroeder M.R., Atal B.S., Hall J.L.* «Optimizing digital speech coders by exploiting masking properties of the human ear», Journal of the Acoustical Society of America, vol.66, pp.1647–1652, 1979.
19. *Петровский А.А.* Объективная оценка качества восстановленного аудио сигнала перцептуальным ПДВП-кодером на базе периферийной модели уха человека // Сборник докладов 5 Международной научной конференции «Цифровая обработка сигналов и её применение» (DSPA'2003), т. 2, Москва, Россия, 2002. С. 123–126.

Павловец Александр Николаевич —

аспирант-заочник в Учреждении образования «Белорусский государственный университет информатики и радиоэлектроники». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Проектирование и технология электронных вычислительных средств». Работает на Заводе вычислительной техники им. С. Орджоникидзе. Область интересов — цифровая обработка речевых сигналов, кодирование речевых сигналов, проектирование проблемно-ориентированных средств вычислительной техники реального времени для мультимедиа-систем.

Петровский Александр Александрович —

доктор технических наук, профессор. Работает в Учреждении образования «Белорусский государственный университет информатики и радиоэлектроники», кафедра «Электронные вычислительные средства». Закончил Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» по специальности «Электронные вычислительные машины». Главные научные интересы лежат в области цифровой обработки сигналов речи и звука для целей компрессии, распознавания, редактирования шума, а также в области проектирования проблемно-ориентированных средств вычислительной техники реального времени для систем мультимедиа.
Член НТО РЭС им. А.С.Попова, IEEE, EURASIP, AES.