# ИСПОЛЬЗОВАНИЕ ЗАКОНОМЕРНОСТЕЙ ПСИХОАКУСТИКИ ПРИ НИЗКОСКОРОСТНОМ КОДИРОВАНИИ РЕЧИ

Серков В.В.[1], Петровский А.А.[2]

Белорусский государственный университет информатики и радиоэлектроники,
ул. П.Бровки 6, Минск, Беларусь, кафедра электронных вычислительных средств
[1]e-mail: sercov@gmx.net [2] e-mail: palex@it.org.by

**Реферат.** В докладе представляется способ использования закономерностей психоакустики при реализации низкоскоростного вокодера с высоким качеством синтезируемой речи. Работа ориентирована на модель речеобразования с раздельным представлением спектров тональной и шумовой компонент речевого сигнала. Психоакустические преобразования производятся на двух этапах: при построении огибающих спектров и при расчете весовых коэффициентов для взвешивания ошибки квантования параметров. Данный подход позволяет уменьшить по сравнению с существующими подходами построения вокодеров ощутимые слушателем искажения.

## 1. Введение

Для кодирования речевой информации с низкими (менее 4,8 кбит/с) скоростями используются вокодеры, модель речеобразования которых основана на параметрическом представлении сигнала. Как правило, одним из параметров модели является кратковременный спектр Фурье речевого сигнала. К спектральным вокодерам относятся, например, Regular Pulse Excitation алгоритм, используемый в системе GSM, MultiBand Excitation, Sinusoidal Transform Coder, различные варианты CELP и пр. [1] Несмотря на значительные успехи в данной области, актуальной является задача дальнейшего снижения скорости передачи и повышения качества синтезированной речи. Большие перспективы в этом имеет по возможности полный учет при кодировании особенностей восприятия звука человеком, изучаемых психоакустикой.

В настоящей работе предлагается новый способ использования закономерностей психоакустики, который позволяет более качественно строить огибающие спектров и рассчитывать весовые коэффициенты для взвешивания ошибки квантования параметров.

## 2. Модель речеобразования

В работе была использована модель речеобразования, основанная на представлении речевого сигнала в виде спектров тональной и шумовой компонент [2]. Данная модель находит все большее применение (напр., [3]). Тональная компонента представляет собой совокупность гармоник основного тона. Каждая гармоника является синусоидой, амплитуда и частота которой линейно изменяются между окнами анализа. Шумовая компонента определяется как разность между исходным сигналом и синтезированной тональной компонентой. Параметрами модели являются частота основного тона, амплитуды гармоник и огибающая спектра шумовой компоненты. Структурная схема кодера, обеспечивающего нахождение данных параметров, показана на рис.1.

При синтезе речи периодическая составляющая создается набором перестраиваемых синусоидальных генераторов. Фазы гармоник основного тона определяются в зависимости от их амплитуд и траектории фундаментальной частоты таким образом, чтобы обеспечить минимальные реверберацию и «металличность». Синтез шумовой компоненты заключается во взвешивании спектра белого шума соответствующей огибающей и обратном преобразовании во временную область. Сформированные периодическая и шумовая составляющие складываются. На рис.2 приведена структурная схема декодера. Подробности процедур анализа и синтеза опубликованы в [4].

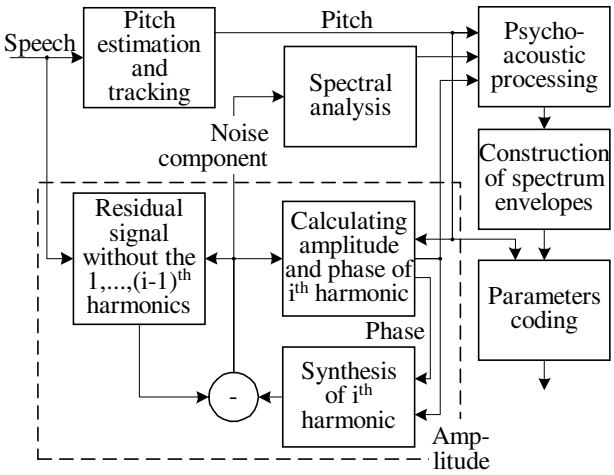


Рис.1. Структурная схема кодера

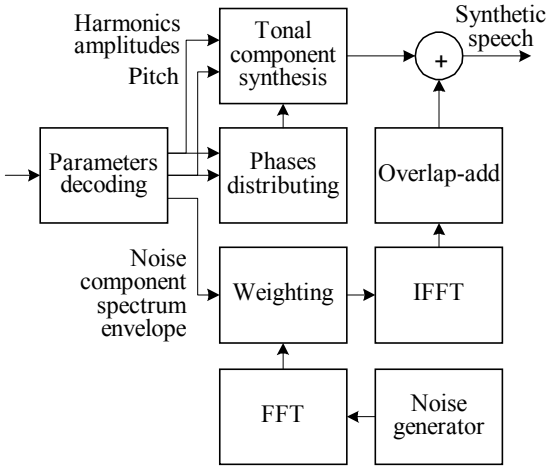


Рис.2. Структурная схема декодера

### 3. Построение огибающих спектров

При построении огибающих спектров ставились четыре основные цели. Спектр речевого сигнала содержит большое количество отсчетов, и, следовательно, не может напрямую использоваться для квантования. Поэтому первой из целей являлось уменьшение количества спектральных отсчетов.

Следующая цель – повышение качества огибающей спектра для речи с высокими частотами основного тона. В используемой чаще всего технике линейного предсказания частотная характеристика фильтра является, по сути, усредненным спектром мощности входного сигнала, а не его огибающей. Это отличие не играет заметной роли в случае невокализованных звуков или для вокализованных звуков с низкой частотой основного тона. При снижении данной частоты частотная характеристика фильтра начинает следовать за «провалами» спектра между отдельными гармониками; искажаются определяемые параметры формант.

Третья цель состоит в учете психоакустических закономерностей. Традиционный способ расчета огибающих спектра заключается в минимизации суммарной квадратичной ошибки в координатах дБ-Гц. При этом одна и та же погрешность аппроксимации в зависимости от амплитуды и частоты воспринимается на слух неодинаково. Для уменьшения слышимых искажений сигнала чаще всего применяют перцептуальные взвешивающие фильтры или регулируют точность квантования в зависимости от порога маскирования. Оба этих подхода не позволяют в полной мере учесть известные закономерности психоакустики. В настоящей работе перед построением огибающих выполнялось преобразование спектральных шкал, что позволило аппроксимировать спектры в линейных для слуха координатах.

И, наконец, последняя цель является специфической для гармонических вокодеров. При квантовании амплитуд гармоник возникают сложности, обусловленные постоянным изменением их количества. Построение огибающей гармонического спектра с фиксированным количеством параметров позволяет устранить данную проблему.

Как известно, спектральное разрешение слухового аппарата человека зависит от частоты: на верхних частотах оно хуже, чем на нижних. Выравнивание спектрального разрешения достигается с помощью преобразования Гц-барк [5]. В соответствии с ним частотная ось трансформировалась следующим образом:

$z = 13*\arctan(0.76*f)+3.5*\arctan((f/7.5)^2)$,

где f и z – частота, выраженная в кГц и барк соответственно.

Слух обладает неодинаковой чувствительностью к энергии на разных частотах. Графически этот факт можно представить в виде кривых равной громкости [6], которые показаны на рис.3. Вдоль каждой кривой уровень громкости, измеряемый в фонах, остается постоянным и полагается равным уровню звукового давления в дБ на частоте 1 кГц. Как видно, кривые семейства схожи между собой, и их можно аппроксимировать абсолютным порогом слышимости, поднимая его по амплитуде. Исходя из этого, компенсация частотной зависимости чувствительности к энергии (преобразование дБ-фон) выполнялось в виде

$P=D-ATH+ATH_{1\ \mathrm{kHz}}$,

где $D$ и $P$ – амплитуды спектральной компоненты в дБ и фонах соответственно; $ATH$ и $ATH_{1\ \mathrm{kHz}}$ – значения абсолютного порога слышимости [5] на частотах данной спектральной компоненты и 1 кГц в дБ:

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

где $f$ – частота в кГц.

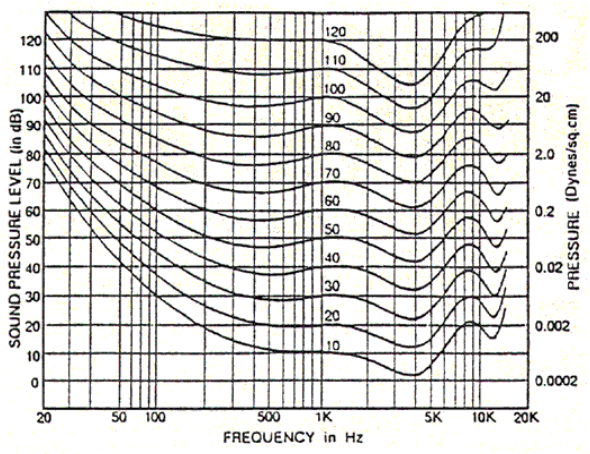


Рис.3. Кривые равной громкости

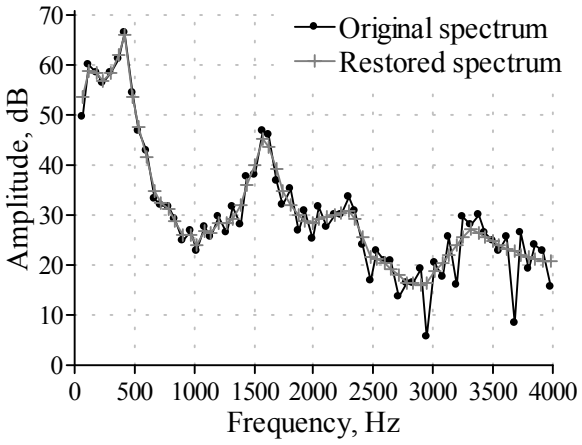


Рис.4. Восстановление спектра по огибающей

Кроме рассмотренных, следует учесть еще один вид нелинейности при восприятии звука: приращение в фонах, требуемое для удвоения субъективной громкости, зависит от уровня громкости. Для устранения данной нелинейности использовалось преобразование фон-сон [7]:

$$S = \begin{cases} 2^{(P-40)/10} & \text{если } P \geq 40 \\ (P/40)^{2.642} & \text{если } P < 40 \end{cases},$$

где S – громкость в сонах.

Огибающие спектров тональной и шумовой компонент сигнала в настоящей работе строились с помощью кусочно-линейной аппроксимации в координатах сон-барк. На рис. 4 приведен пример восстановления тонального спектра по его огибающей. Использование психоакустических шкал способствует более точной аппроксимации в области нижних частот. Исследования показали, что для тональной составляющей речи уменьшение количества спектральных отсчетов в огибающей на 30% не заметно на слух. В случае шумовой компоненты это количество может быть еще уменьшено.

### 4. Расчет весовых коэффициентов для взвешивания ошибок квантования

В качестве весовых коэффициентов при взвешивании ошибок квантования использовались пороги частотного маскирования. В отличие от широко распространенной психоакустической модели MPEG [8], разделение сигнала на тональную и шумовую компоненты позволяет обойтись без определения степени тональности. При этом известные правила маскирования для случаев тон-маскирует-шум и шум-маскирует-тон [5] применялись в «чистом» виде: порог маскирования, рассчитанный по спектру тональной компоненты, использовался для квантования шумовой и наоборот. Другое отличие состояло в том, что тональные маскеры (гармоники основного тона) не группировались в пределах критических полос, а учитывались раздельно. За счет этого удалось повысить точность для средних и высоких голосов.

Пороги маскирования вычислялись следующим образом:

$$MT_i = ATH_i + 10^{\log_{10}\left(\sum_{j=1}^{N} SF(\Delta f)*P_j\right) - O_i/10},$$

где $ATH_i$ – значение абсолютного порога слышимости на частоте $i$-й компоненты спектра в разах; $SF$ – функция распространения возбуждения по базилярной мембране [9]; $\Delta f$ – разность частот $i$-й компоненты спектра и $j$-го маскера в барках; $P_j$ – мощность $j$-го маскера в разах; $O_i$ – смещение маскирования [10].

Функция распространения определялась как

$$SF(x) = 10^{15.81+7.5(x+0.474)-17.5\sqrt{1+(x+0.474)^2}/10}.$$

Смещение маскирования:

$$O_i = \begin{cases} 14.5 + z_i & \text{для случая тто - маскирует - шум} \\ 2.0 + 2.05\arctan\left(f_i/4\right) - 0.75\arctan\left(f_i^2/2.56^2\right) & \text{для случая шум - маскирует - тон} \end{cases},$$

где $z_i$ и $f_i$ – частоты $i$-й компоненты спектра в барках и кГц соответственно.

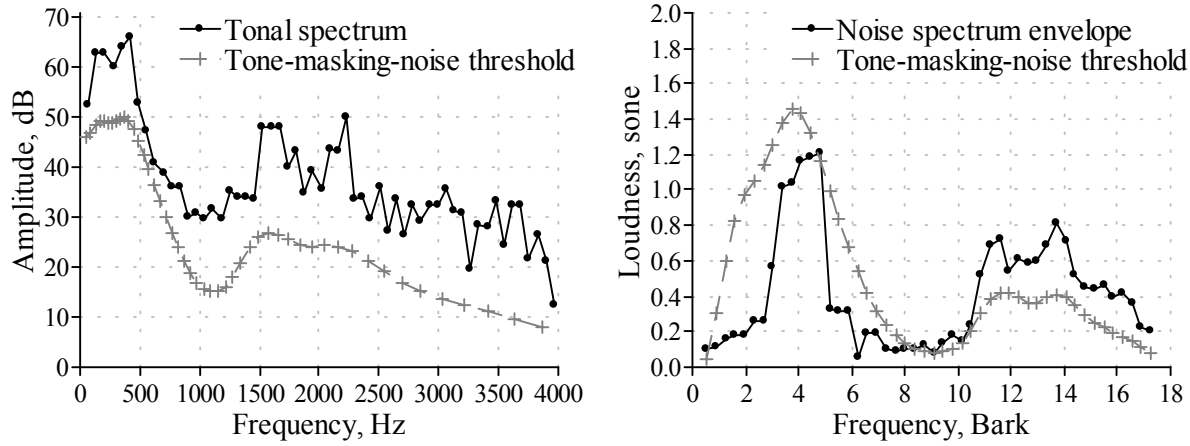


Рис.5. Пример психоакустических преобразований для случая тон-маскирует-шум

### 5. Заключение

Приведенный способ учета психоакустических закономерностей позволяет повысить точность распределения шумов квантования для вокодеров, использующих модель речеобразования тон+шум [2]. На рис. 5 показан пример психоакустических преобразований для случая тон-маскирует-шум. При тестировании психоакустической модели рассчитанные пороги использовались для маскирования попавших под них спектральных отсчетов. Для приведенного примера низкочастотная часть спектра шумовой компоненты

полагалась равной нулю. После этого речевой сигнал заново синтезировался. Эксперименты показали, что заметной на слух деградации качества синтезированной речи не возникает.

В настоящее время ведется разработка квантователя, способного эффективно учитывать получаемые весовые коэффициенты при взвешивании ошибок квантования. Он реализуется в виде искусственной нейронной сети на основе многослойного автоассоциативного перцептрона.

Литература

1. A.M.Kondoz, "Digital Speech: Coding for Low Bit Rate Communication Systems", New York, NY: John Wiley & Sons, 1996.
2. V.Sercov, A.Petrovsky, "An Improved Speech Model with Allowance for Time-Varying Pitch Harmonic Amplitudes and Frequencies in Low Bit-Rate MBE Coders", *//proc. of the 6ht European Conf. on Speech Communication and Technology*, EUROSPEECH'99, Budapest, Hungary, 1999, pp. 1479-1482.
3. Ç.Ö.Etemoğlu, V.Cuperman, A.Gersho "Speech Coding with an Analysis-by-Synthesis Sinusoidal Model", *// proc. of the IEEE Int. Conf. Acoust. Speech Signal Proc.*, ICASSP`2000, vol. III, Istanbul, 2000, pp. 1371-1374.
4. A.Petrovsky, V.Sercov, "Low Bit-Rate AbS Spectral Coding Based on the Harmonic Analysis of Speech Agreed upon with Time-Varying Pitch Frequency and Psychoacoustical Optimization", *// proc. of the Nordic Signal Proc. Symp.*, NORSIG'2000, Sweden, 2000, pp.45-48.
5. E.Zwicker, H.Fastl, "Psychoacoustics: Facts and Models". Berlin: Springer-Verlag, 1990.
6. D.Robinson, R.Dadson, "A Redetermination of the Equal-Loudness Relations for Pure Tones", *// Bri. J. Appl. Physics*, 1956, pp.166-181.
7. R.Bladon, "Modeling the Judgement of Vowel Quality Differences", *// J. Acoust. Soc. Amer.,* vol. 69, 1981, pp. 1414-1422.
8. ISO/IEC 11172-3, "Information Technology – Coding of, part 3: Audio", ISO/IEC, 1993.
9. M.Schroeder, B.Atal, L.Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear", *// J. Acoust. Soc. Amer.,* vol. 66, 1979, pp. 1647-1652.
10. R.Kapust, "A Human Ear Related Objective Measurement Technique Yields Audible Error and Error Margin", *// proc. 11th Int. AES Conf.,*Portland, OR, 1992, pp. 191-202.

## APPLYING PSYCOACOUSTICS IN LOW BIT-RATE TONE+NOISE SPECTRAL SPEECH CODING

Sercov, V.[1], Petrovsky, A.[2]
Department of Computer Engineering, Belarusian State Univ. of Informatics and Radioelectronics,
6, P.Brovky Str., 220027, Minsk, BELARUS
[1]e-mail: sercov@gmx.net [2]e-mail: palex@it.org.by

### 1. Introduction

Spectral vocoders are normally used for low bit-rate coding (less than 4,8 кbps) [1]. Further bit-rate reducing and improving the quality of the synthesyzed speech can be achieved by extended employing the methods of psychoacoustics. The paper suggests new method of using the psychoacoustic laws, which creates better spectrum envelopes and allows precise calculation of the coefficients for weighting parameter quantization errors.

### 2. Building spectral envelopes

The work employs the model of speech encoding based on tonal plus noise spectral representation of the speech signal [2]. When building envelopes of the spectrum the following aims were set: reducing the number of the spectrum samples, improving the quality of the envelope for high-pitched speech, taking into account pscychoacoustic laws, describing the harmonic spectrum by fixed set of parameters.

It is known that human ear is not equally sensitive to the different frequencies. The Bark-Herz transform was employed in order to level the spectral resolution the frequency axis [5]:

$z = 13*\arctan(0.76*f)+3.5*\arctan((f/7.5)^2)$,

where f and z – frequency in kHz and Bark respectively.

Human ear percepts energy at various frequencies differently. Along each of the equal loudness level curves [6] the loudness level expressed in phones remains unchanged and is equal to sound pressure level in dB at 1kHz. The curves poses the same form and can be be approximated by absolute threshold of hearing by increasing its amplitude. Thus, the compensation of the energy sensitivity at different frequencies (dB to phon conversion) was implemented in the following form:

$P=D-ATH+ATH_{1\ \mathrm{kHz}}$,

where $D$ and $P$ – amplitudes of the spectral component in dBs and phones respectively; $ATH$ and $ATH_{1\ \mathrm{kHz}}$ – absolute threshold of hearing values [5] at component's frequencies and at 1kHz in dB:

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

where $f$ – frequency in kHz.

Since doubling of the subjective loudnes requires different phone increase depending on the loudness level, the following phon to sone conversion is used [7]:

$$S = \begin{cases} 2^{(P-40)/10} & \text{если } P \geq 40 \\ (P/40)^{2.642} & \text{если } P < 40 \end{cases},$$

where S – loudness in sones.

Spectrum envelopes of the tonal and noise components in this work are piecewise-linear approximaded in sone-Bark coordinates. Using psychoacoustic scales enables better approximation in low frequency domain. The research showed that 30% reducing of number of the envelope samples is not hearable. In case of the noise component this number can be reduced even to greater extent.

### 3. Calculating weight coefficients for quantization errors

Weighting the quantization errors was performed through the use of frequency masking thresholds. In contrast to widely used MPEG model [8], discriminating the signal on tonal and noise components eliminates the necessity of determining the voicing degree. In this case the known "tone masking noise" and "noise masking tone" were used in their "pure" form: masking threshold calculated by tonal component's spectrum was employed for quantization of the noise part and vise versa. There is another difference connected with tonal maskers (pitch harmonics). They were not grouped within critical bands. It made possible the quality improvement in middle- and high pitched voices.

The masking thresholds were calculated according to the equation:

$$MT_i = ATH_i + 10^{\log_{10}\left(\sum_{j=1}^{N} SF(\Delta f)*P_j\right) - O_i/10},$$

where $ATH_i$ – absolute threshold of hearing at $i$-th specrum component's frequency; $SF$ – spreading function of exitation along basilar membrane [9]; $\Delta f$ – frequency difference between $i$-th spectrum component and $j$-th masker in Barks; $P_j$ – power of $j$-th masker; $O_i$ – masking offset [10].

Spreading function was calculated as:

$$SF(x) = 10^{15.81+7.5(x+0.474)-17.5\sqrt{1+(x+0.474)^2}/10}$$

and masking offset:

$$O_i = \begin{cases} 14.5 + z_i & \text{tone masking noise case} \\ 2.0 + 2.05\arctan\left(f_i/4\right) - 0.75\arctan\left(f_i^2/2.56^2\right) & \text{noise masking tone case} \end{cases},$$

where $z_i$ and $f_i$ – frequencies of the $i$-th spectrum component in Barks and kHz respectively.

### 4. Conclusion

The suggested method of employing the psychoacoustic laws increases the pricision of noise distribution for vocoders that are based on tone+noise speech model [2]. When testing the psychoacoustic model, the calculated thresholds where used for masking the spectrum samples with lower amplitudes. After that the speech signal was reconstructed. The experiments showed that the masking does not lead to percepted degradation of speech quality.

The research aimed at creating the quantatizator able of efficient employing the coefficients received by weighting the quantatization errors is now conducted. The quantatizator is implemented through artificial neural network built on multilayer autoassociative perceptron.

References

1. A.M.Kondoz, "Digital Speech: Coding for Low Bit Rate Communication Systems", New York, NY: John Wiley & Sons, 1996.

2. V.Sercov, A.Petrovsky, "An Improved Speech Model with Allowance for Time-Varying Pitch Harmonic Amplitudes and Frequencies in Low Bit-Rate MBE Coders", *//proc. of the 6ht European Conf. on Speech Communication and Technology*, EUROSPEECH'99, Budapest, Hungary, 1999, pp. 1479-1482.

3. E.Zwicker, H.Fastl, "Psychoacoustics: Facts and Models". Berlin: Springer-Verlag, 1990.

4. D.Robinson, R.Dadson, "A Redetermination of the Equal-Loudness Relations for Pure Tones", *// Bri. J. Appl. Physics*, 1956, pp.166-181.

5. R.Bladon, "Modeling the Judgement of Vowel Quality Differences", *// J. Acoust. Soc. Amer.,* vol. 69, 1981, pp. 1414-1422.

6. ISO/IEC 11172-3, "Information Technology – Coding of, part 3: Audio", ISO/IEC, 1993.

7. M.Schroeder, B.Atal, L.Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear", *// J. Acoust. Soc. Amer.,* vol. 66, 1979, pp. 1647-1652.

8. R.Kapust, "A Human Ear Related Objective Measurement Technique Yields Audible Error and Error Margin", *// proc. 11th Int. AES Conf.,*Portland, OR, 1992, pp. 191-202.