MSU Audio Codecs Comparison In Time And Time-Frequency Domain

Measures in time and time-frequency domains

Руководитель проекта: Александр Жирков
Замеры: Валентин Вербовой
Обработка: Александр Страбыкин
Текст: Валентин Вербовой, Александр Страбыкин
Консультант: Дмитрий Ватолин

Произведено подробное сравнение десяти распространенных кодеков на битрейтах 8-192 kbps несколькими методиками PSNR (power signal-to-noise ratio) - тестирования

September 2003

CS MSU Graphics & Media Lab

Audio Group

http://www.compression.ru/audio/

Сравнительное PSNR-тестирование звуковых кодеков

1. Задачи исследования

Основной задачей исследования являлось определение эффективной метрики для сравнения качества кодирования звука различными аудио кодеками. Основной упор был сделан на сравнение качества кодирования специализированных для речи кодеков (вокодеров) и универсальных кодеков.

В рамках тестирования были использованы следующие кодеки:

Речевые кодеки:

1) GSM 6.10 (встроенный кодек windows 98 SE), **bitrate 16,32 и 72 kbps;**
2) CELP (встроенный кодек windows 98 SE), **bitrate 8 kbps;**
3) TrueSpeech (встроенный кодек windows 98 SE), **bitrate 8 kbps.**

Универсальные кодеки:

1) MP3 кодек Lame 3.93 MMX (RazorLame V 1.1.5.1342), **bitrate 8,32,64,96,192 kbps;**
2) MP3 кодек из пакета Blaze Media Convert 1.4 (BMC), **bitrate 8, 32, 64, 96, 192 kbps;**
3) MP3 Pro Fraunhofer (CoolEdit 2.0 mp3pro audio coding technology licensed from Coding Technologies, Fraunhofer IIS and Thomson multimedia), **bitrate 32 kbps;**
4) Advanced Audio Coding (AAC) (производитель неизвестен), **bitrate 32, 64, 96, 192 kbps;**
5) Windows Media Audio (WMA) кодек из пакета Blaze Media Convert 1.4, **bitrate 32, 64, 96, 192 kbps;**
6) Yamaha Sound VQ Format (VQF), **bitrate approx. 78kbps.**

2. Методика тестирования

2.1 Тестовые файлы

Для тестирования кодеков было выбрано 5 тестовых файлов:

Speech.wav – этот файл содержит записанный фрагмент программы радио-новостей. Он был добавлен в набор, чтобы оценить предполагаемый выигрыш в качестве и размерах файла, получаемый при использовании вокодеров при кодировании речи.

Instrvoice.wav – файл содержит как звуки отдельных инструментов, так и фрагмент пения без музыкального сопровождения.

Music.wav – файл содержит типичный музыкальный фрагмент: песня с аккомпанементом.

Naturenoises.wav – этот файл содержит звуки природы (шум волн, крики птиц) и был добавлен в набор для того, чтобы выяснить качество кодирования нестандартной звуковой информации.

Test.wav – файл содержит синтетический тест, сонограмма которого представлена на рис. 1.

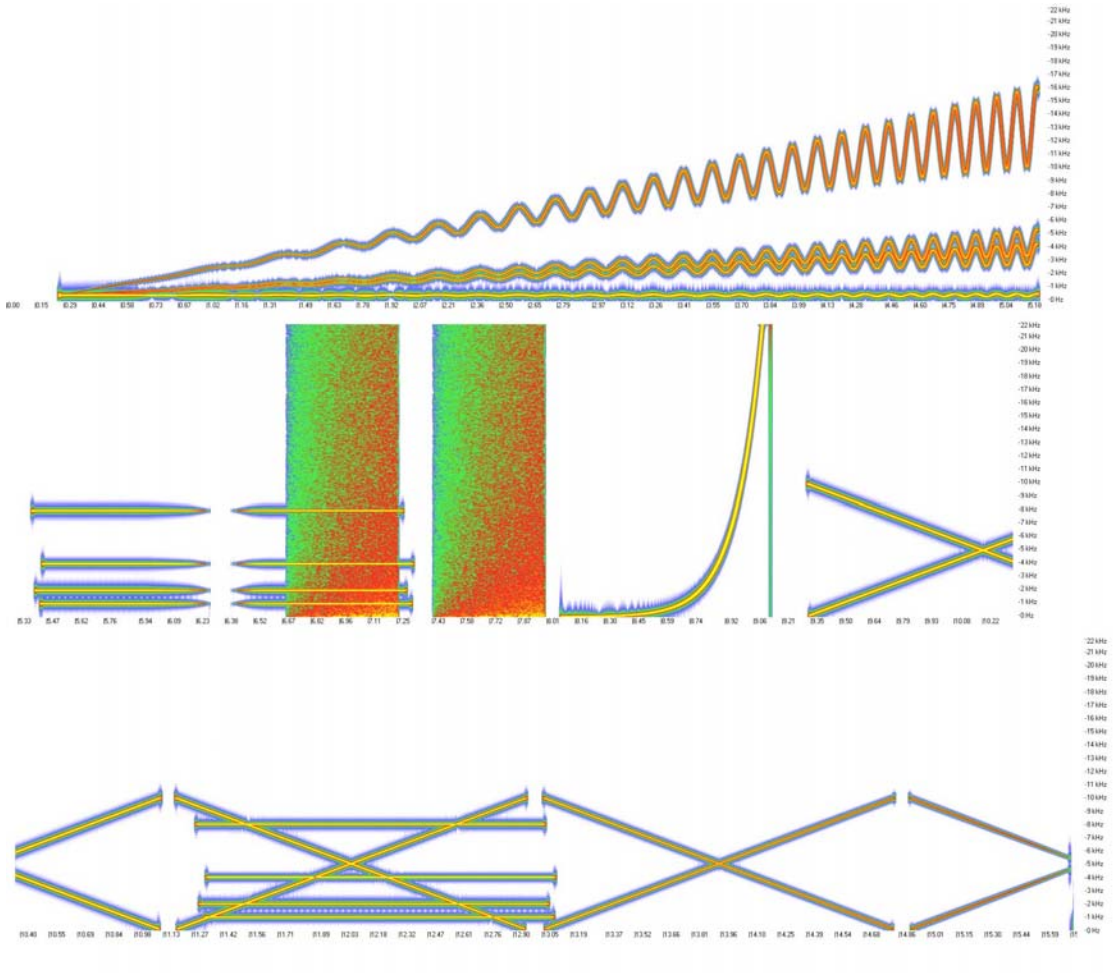


рис. 2.1.1. Сонограмма файла test.wav

Все тестовые файлы содержали шестнадцатибитный монозвук с частотой дискретизации 44100 Гц.

2.2 Используемые метрики

Широкомасштабное тестирование проводилось с использованием двух основных метрик: временной и частотной Peak-Signal-to-Noise-Ratio (PSNR). Общая методика тестирования включала цикл компрессии-декомпрессии исходного файла исследуемыми кодеками с последующим сравнением сжатого и исходного файлов программами реализующими метрики.

2.2.1 Временная PSNR-метрика.

Значение временного PSNR для двух файлов, представленных массивами сэмплов (a[i], i = 1..n; b[j], j = 1..m) вычислялось по следующей формуле:

$$PSNR = 20 \times Log\left(\frac{Max\_sample\_value}{\sqrt{\frac{\sum_{i=1}^{\min(n,m)}(a[i]-b[i])^2}{\min(n,m)}}}\right),$$

где Max_sample_value – максимальное значение амплитуды каждого сэмпла, в нашем случае равное 2^{15} = 32768.

Часть тестируемых кодеков смещала звуковую информацию на некоторый временной интервал, добавляя тишину в начале сигнала или, наоборот, удаляя первые несколько сэмплов. Эксперименты показали, что смещение доходит до 5000 сэмплов, что составляет примерно одну десятую секунды. Смещение такого рода определялось специальным алгоритмом и существенно улучшало адекватность сравнения по PSNR. Более того, эта метрика была расширена, и, кроме PSNR для тестируемых файлов также вычислялись максимальное и среднее отклонения, средние интегральные значения обоих файлов и разница между ними, но в виду их малой информативности в условиях поставленной задачи, в данном тестировании они не использовались.

2.2.2. Частотно-временная PSNR-метрика

Основная часть кодеков перед сжатием переводит сигнал в частотно-временное пространство, поэтому для них более корректно использовать метрику, работающую как с временными составляющими сигнала, так и с частотными.
Основная методика – сравнение отклонения амплитуд сигналов в частотно-временном пространстве. От исследуемых сигналов берется дискретное оконное преобразование Фурье с некоторым шагом по времени STFT (short-time fourier transform), после чего вычисляются амплитуды по формуле:

$$A_n[j] = \log\left(\sqrt{FFT[j].\mathrm{Im}^2 + FFT[j].\mathrm{Re}^2}\right)$$

Каждый такой вектор является спектром сигнала небольшой области вокруг данной временной точки. Объединив все векторы A_n моментальных спектров в столбцы матрицы, можно получить сонограмму данного сигнала.

$$Sono[i][j] = A_i[j]$$

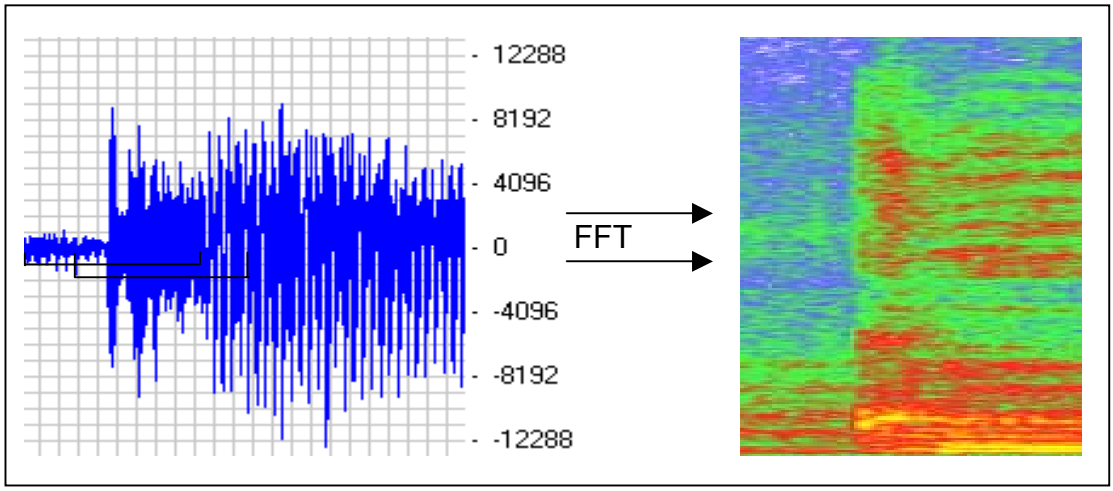


рис 2.2.2 (преобразование сигнала из амплитудно-временного пространства в амплитудно-частотно-временное пространство)

Частотно-временное PSNR для двух сигналов рассчитывалось следующим образом:

$$PSNR = 20 \times Log\left(\frac{Max\_value}{\sqrt{\frac{\sum_{i=0}^{\min(T_1,T_2)} \sum_{j=f_1}^{f_2} (Sono_1[i][j] - Sono_2[i][j])^2}{\min(T_1, T_2)}}} \right)$$

, где

Max_value – амплитуда сигнала максимально возможной мощности, допустимая в данном представлении звукового сигнала; T_1 и T_2 количество векторов моментальных пиков для первого и второго сигнала соответственно; f_1 и f_2 – параметры, отфильтровывающие из общей спектрограммы частотную полосу для исследования. Преимущество данного метода заключается в том, что он не чувствителен к фазе сигнала, в результате чего подходит для кодеков, сохраняющих общее звучание, но не сохраняющих фазы (форму волны), а также в том, что появляется возможность рассмотрения искажений в отдельных частотных диапазонах. Тем не менее, данный метод тестирования нужно применять с большой осторожностью, т.к. он крайне чувствителен к изменению общей амплитуды сигналов. Например, возьмем три сигнала: первый – исходный, второй – полученный из исходного путем добавления слабых щелчков, третий – полученный из исходного путем увеличения его амплитуды на 0.5 Дб. Частотно-временная PSNR метрика выдаст результат о том что сигнал, с чуть большей средней амплитудой будет дальше от оригинала, чем сигнал с наложенными щелчками, хотя на слух он не будет отличаться от исходного, а посторонние звуки (щелчки) человек будет хорошо слышать.

В тестировании рассматриваются три частотных полосы – для низких частот, средних и высоких.

2.3 Дополнительные методы визуализации

Для контроля правильности работы PSNR метрики были использованы следующие простейшие методы визуализации спектральных различий в сигналах: построение “*разницы спектров”*, “*спектра разницы”*, *“разницы сонограмм”* сигналов. Такие методы не дают однозначной автоматической оценки, но во многих случаях помогают понять, с чем связаны результаты, выдаваемые PSNR метриками.

3. Результаты тестирования PSNR-метриками

На следующих диаграммах представлены графики сводных результатов всех тестируемых кодеков, с различными битрейтами. Для каждого из тестируемых файлов было приведено 4 диаграммы: первая построена по данным временной PSNR метрики, а остальные три – частотно-временной PSNR метрики в низкочастотном, среднечастотном и высокочастотном диапазонах.

Диаграмма 1.1

PSNR файла test.wav

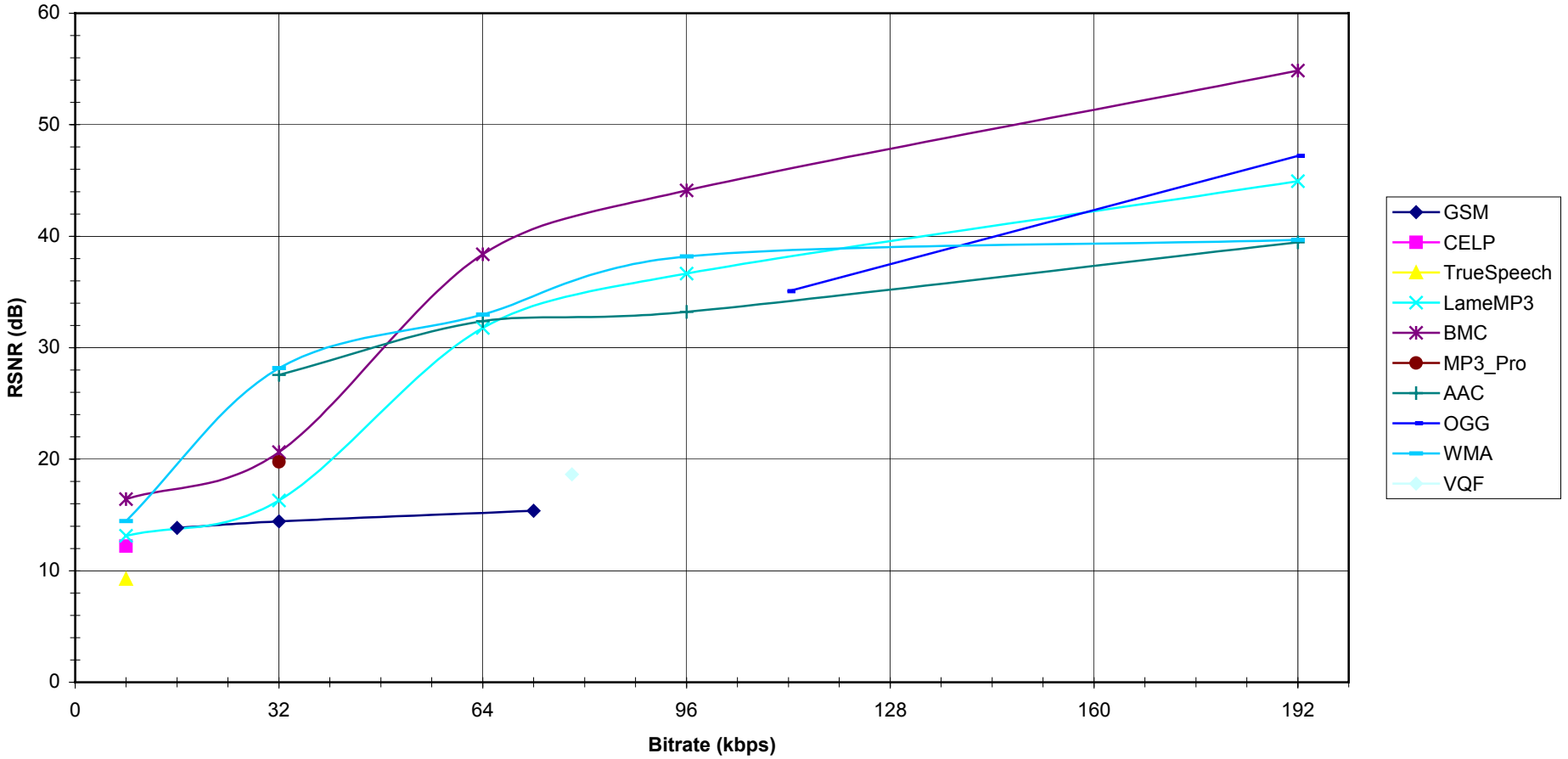


Диаграмма 1.2

PSNR высоких частот файла test.wav

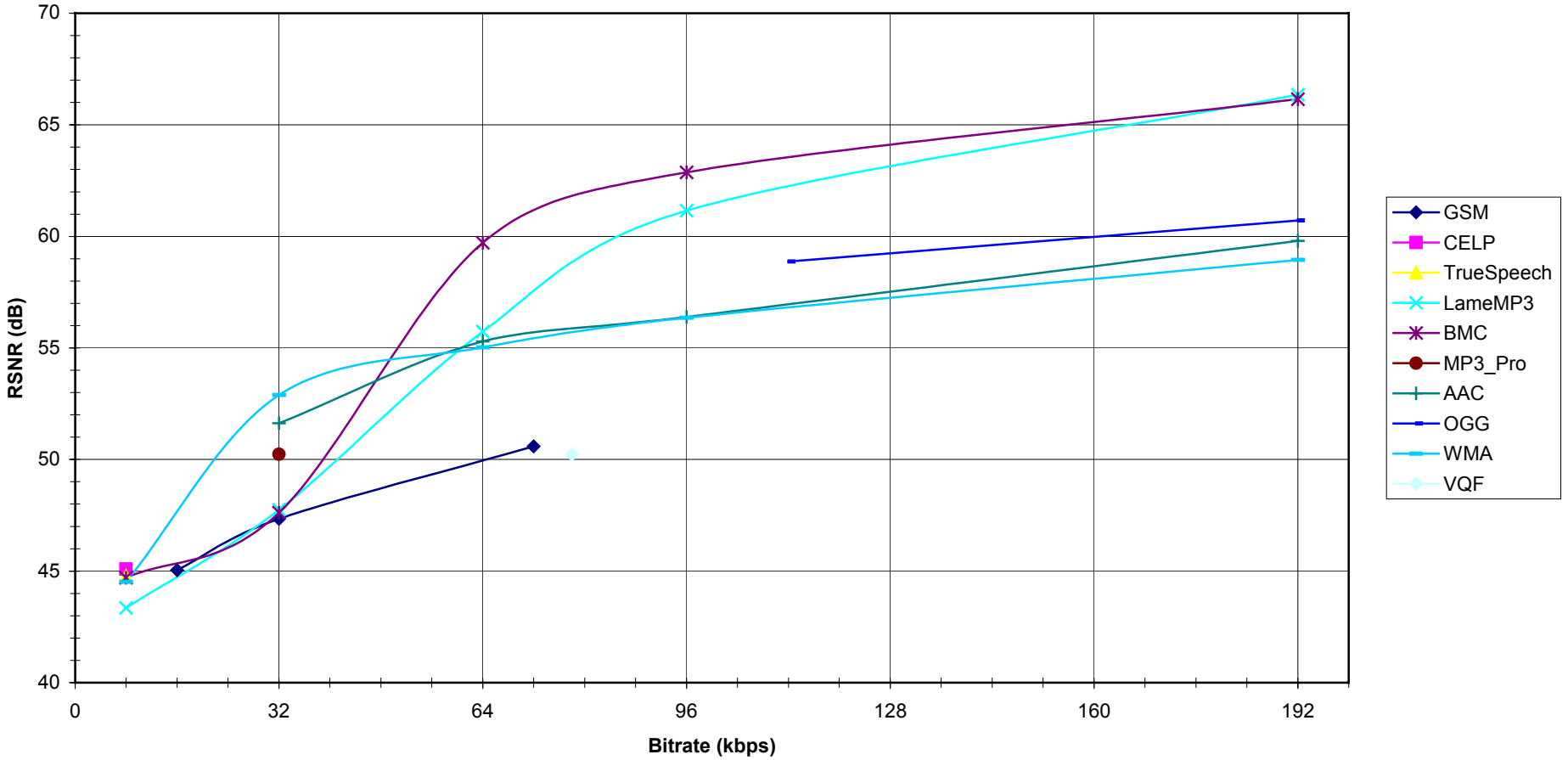


Диаграмма 1.3

PSNR средних частот файла test.wav

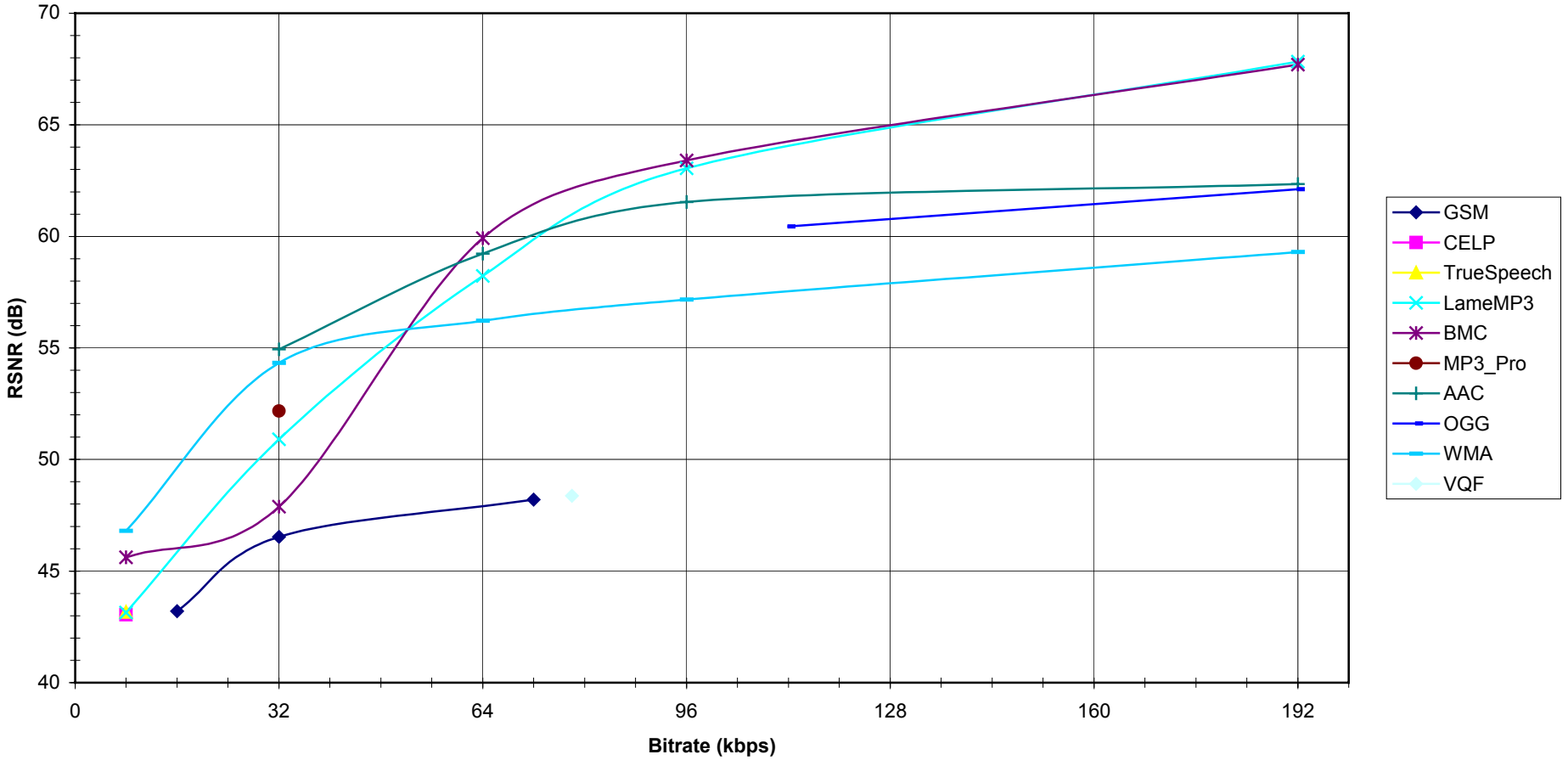


Диаграмма 1.4

PSNR низких частот файла test.wav

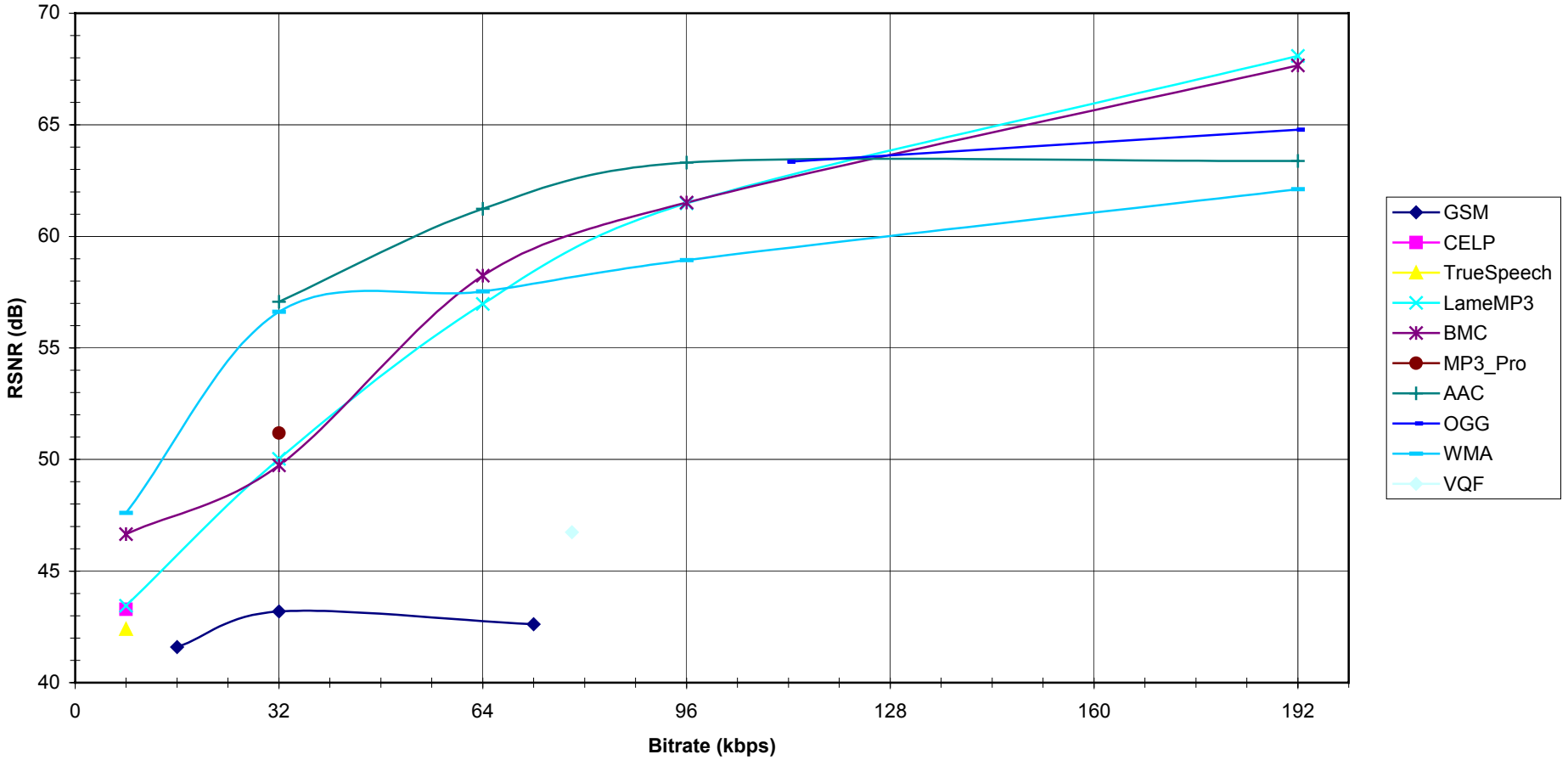


Диаграмма 2.1

PSNR файла speech.wav

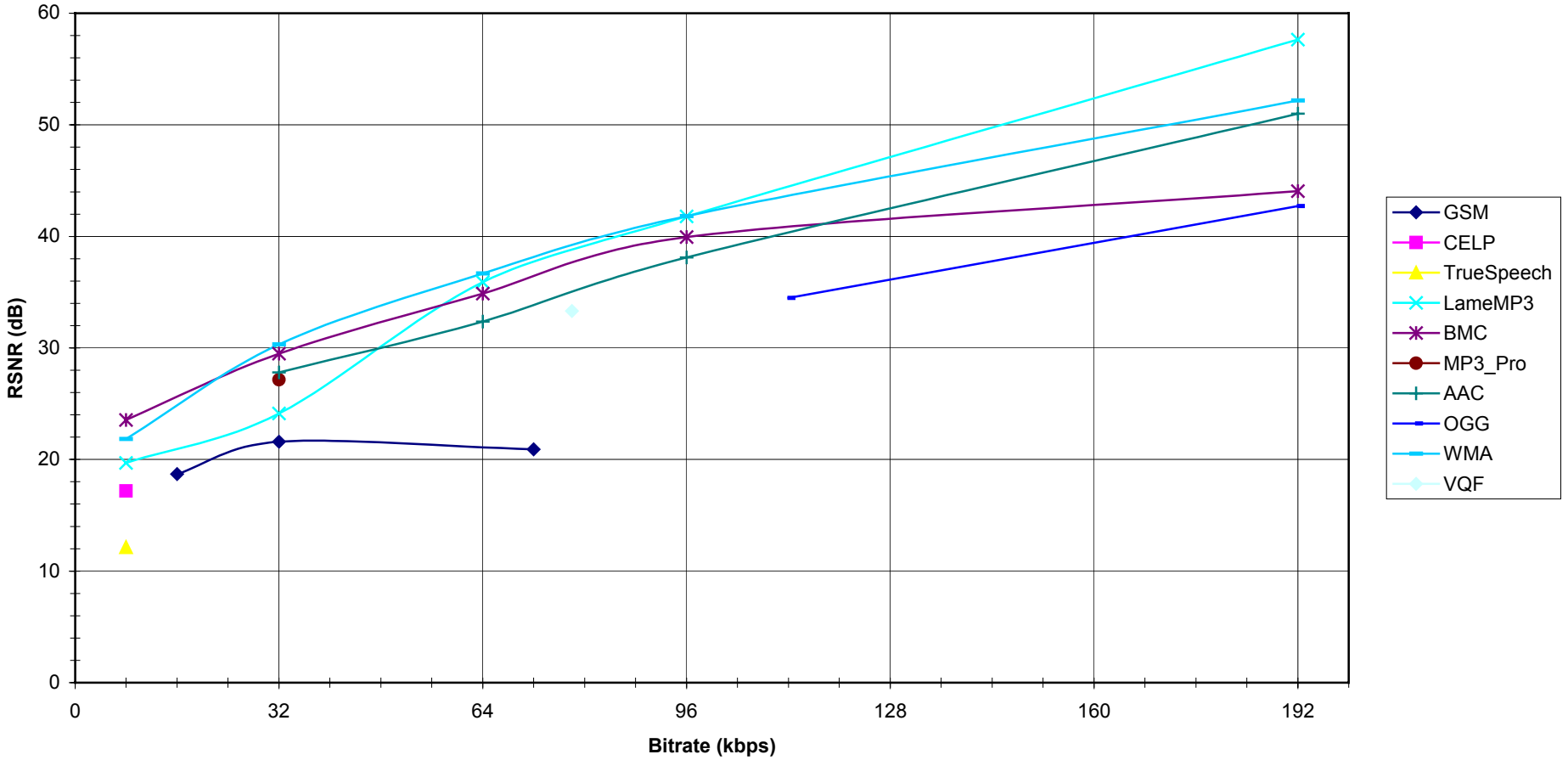


Диаграмма 2.2

PSNR высоких частот файла speech.wav

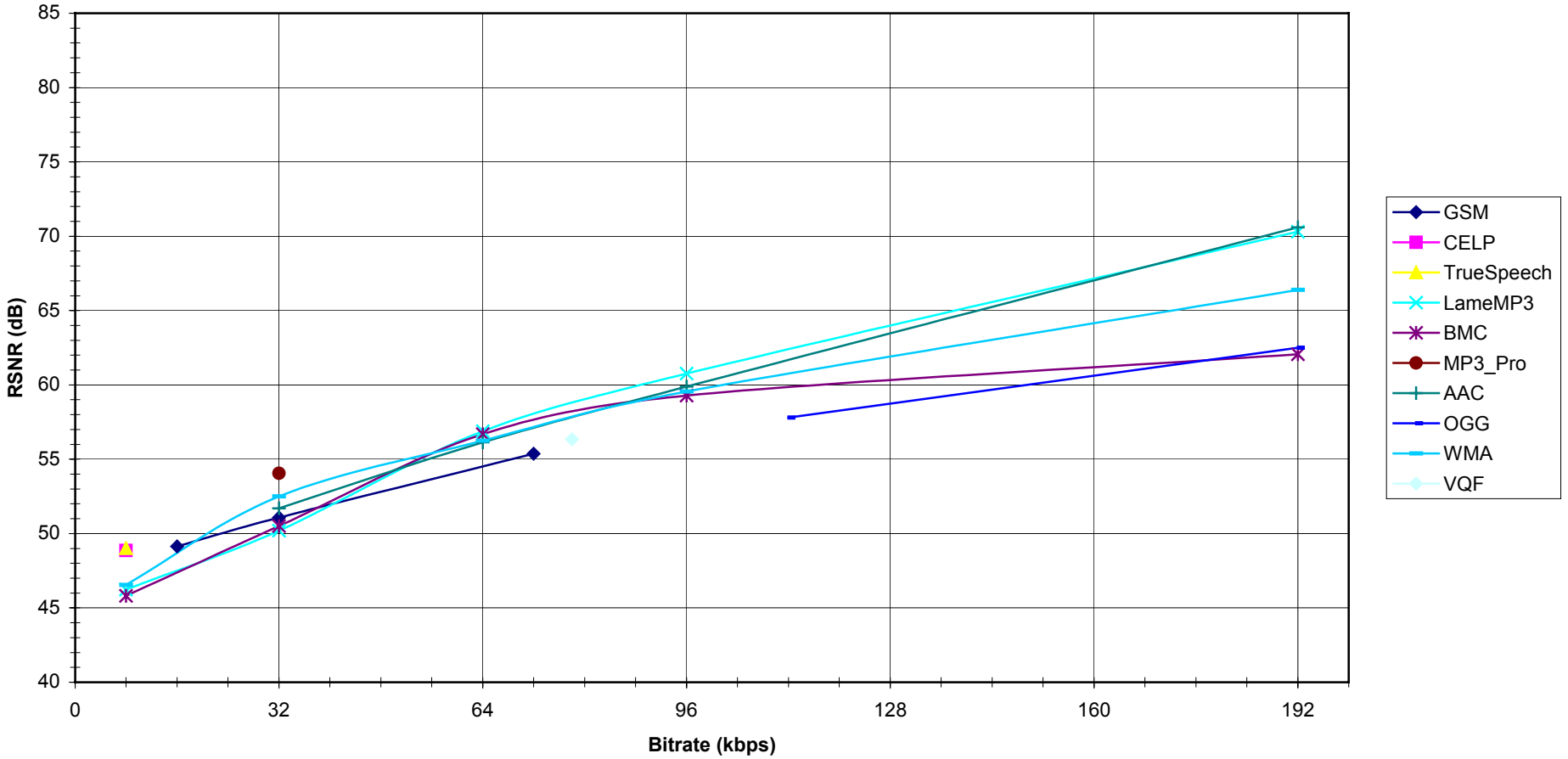


Диаграмма 2.3

PSNR средних частот файла speech.wav

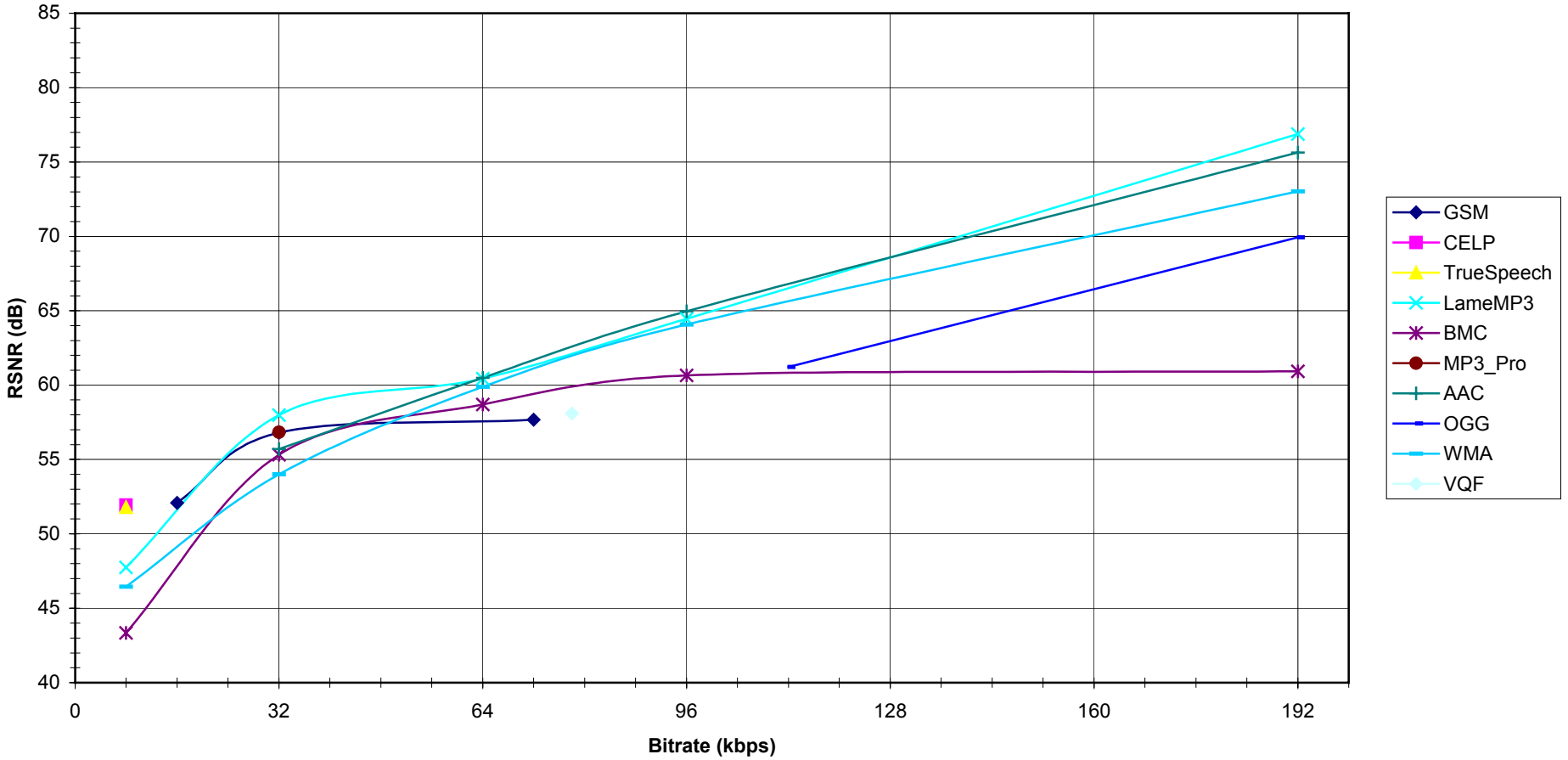


Диаграмма 2.4

PSNR низких частот файла speech.wav

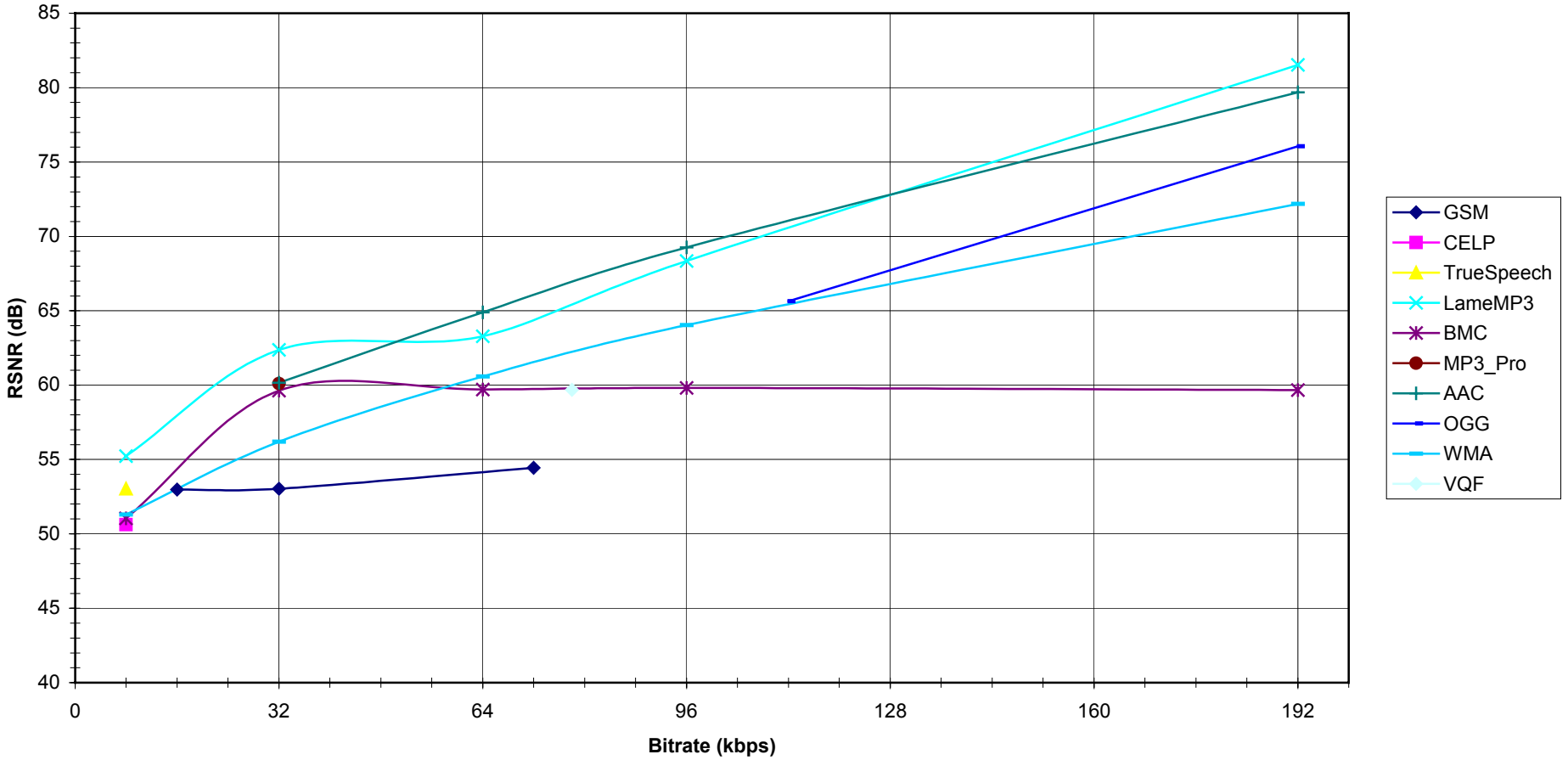


Диаграмма 3.1

PSNR файла naturenoises.wav

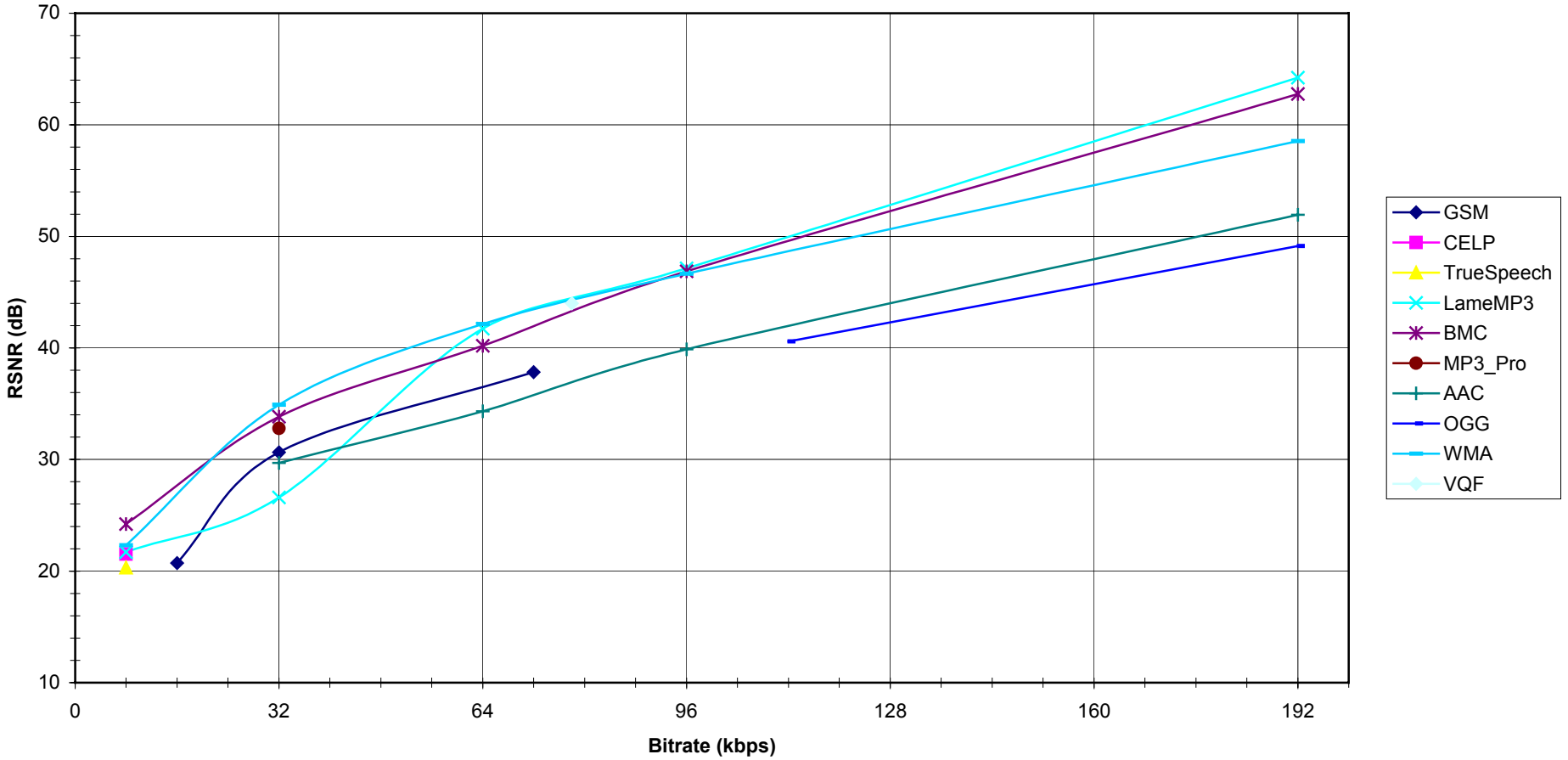


Диаграмма 3.2

PSNR высоких частот файла naturenoises.wav

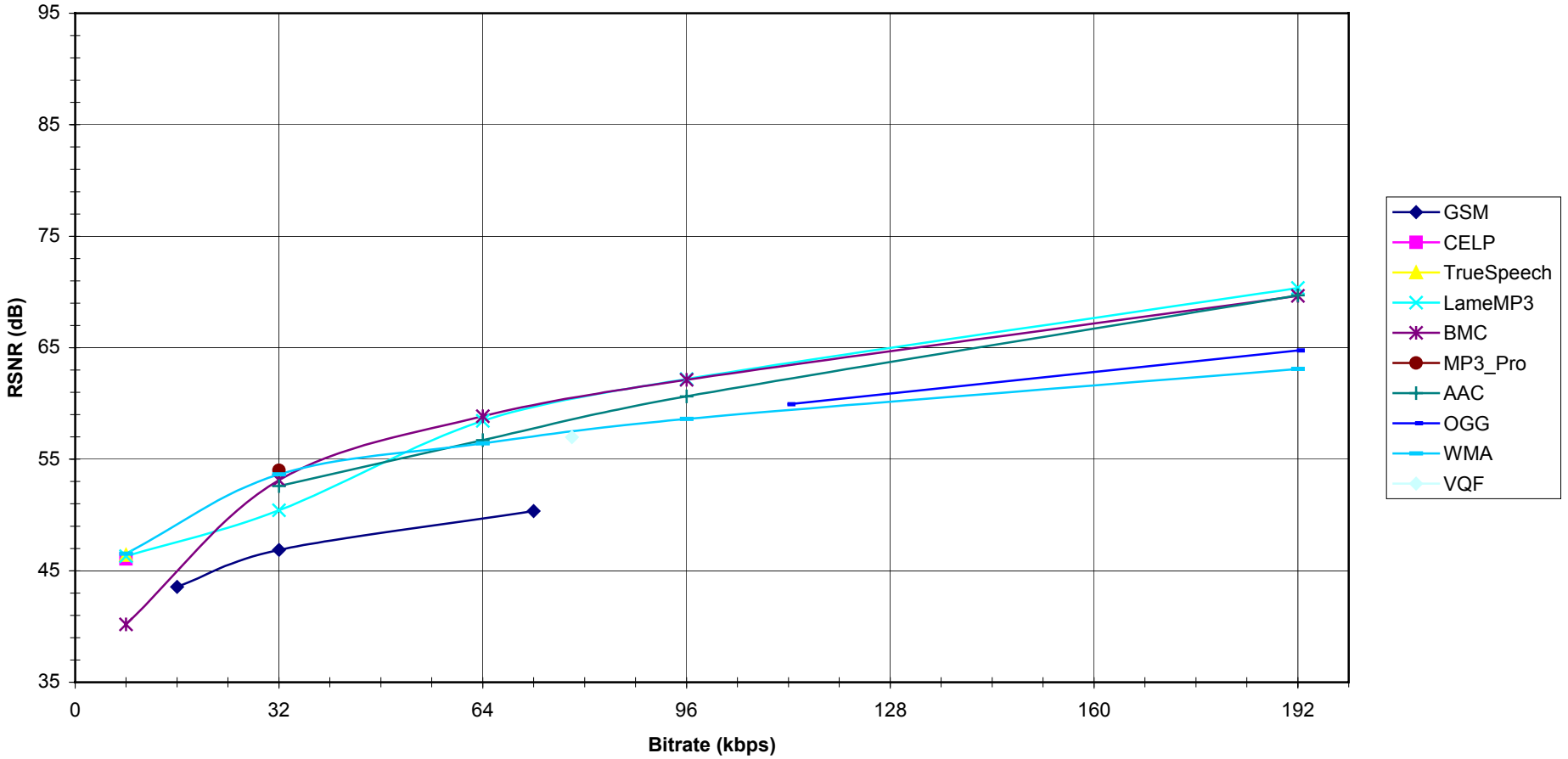


Диаграмма 3.3

PSNR средних частот файла naturenoises.wav

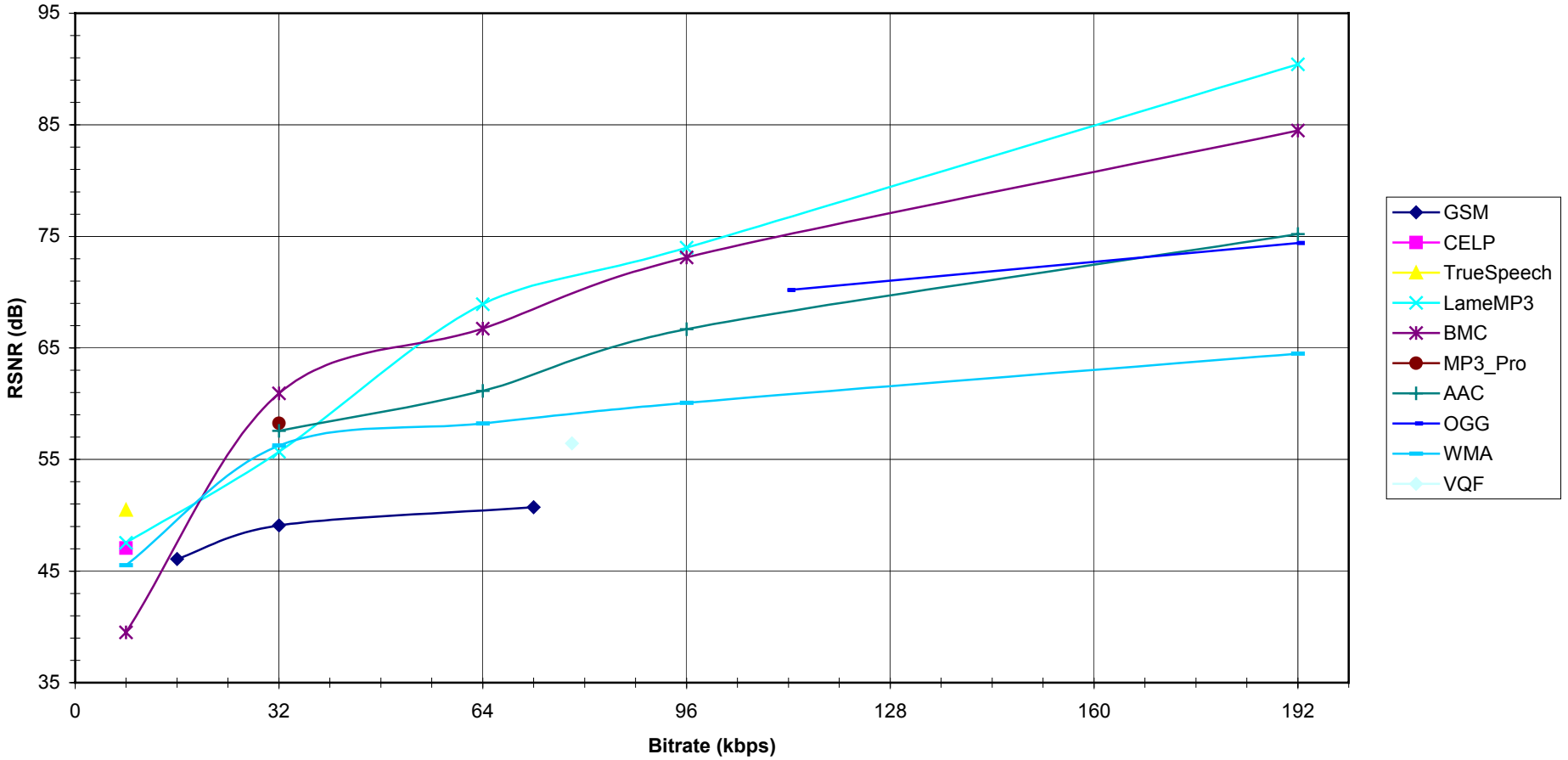


Диаграмма 3.4

PSNR низких частот файла naturenoises.wav

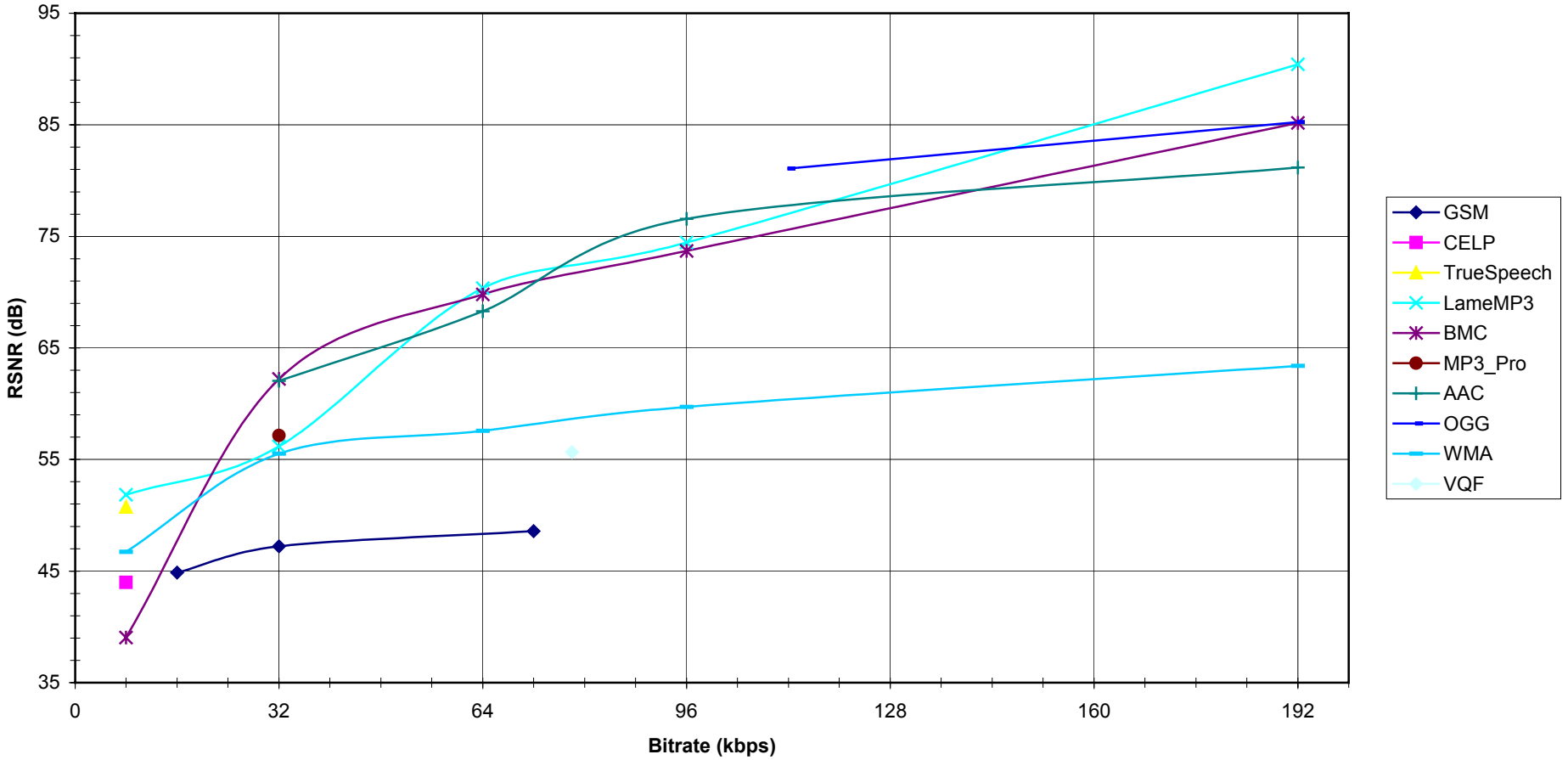


Диаграмма 4.1

PSNR файла music.wav

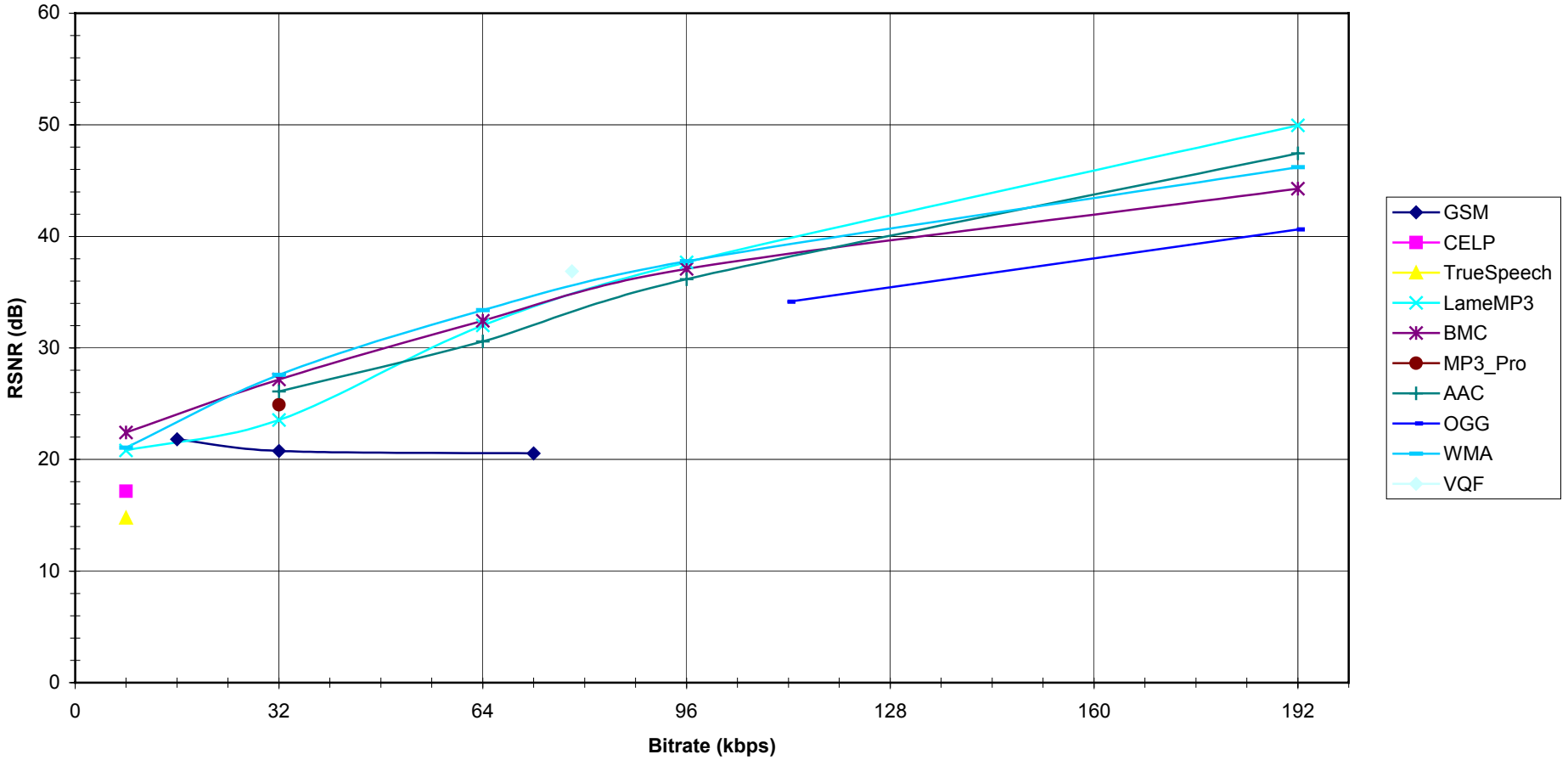


Диаграмма 4.2

PSNR высоких частот файла music.wav

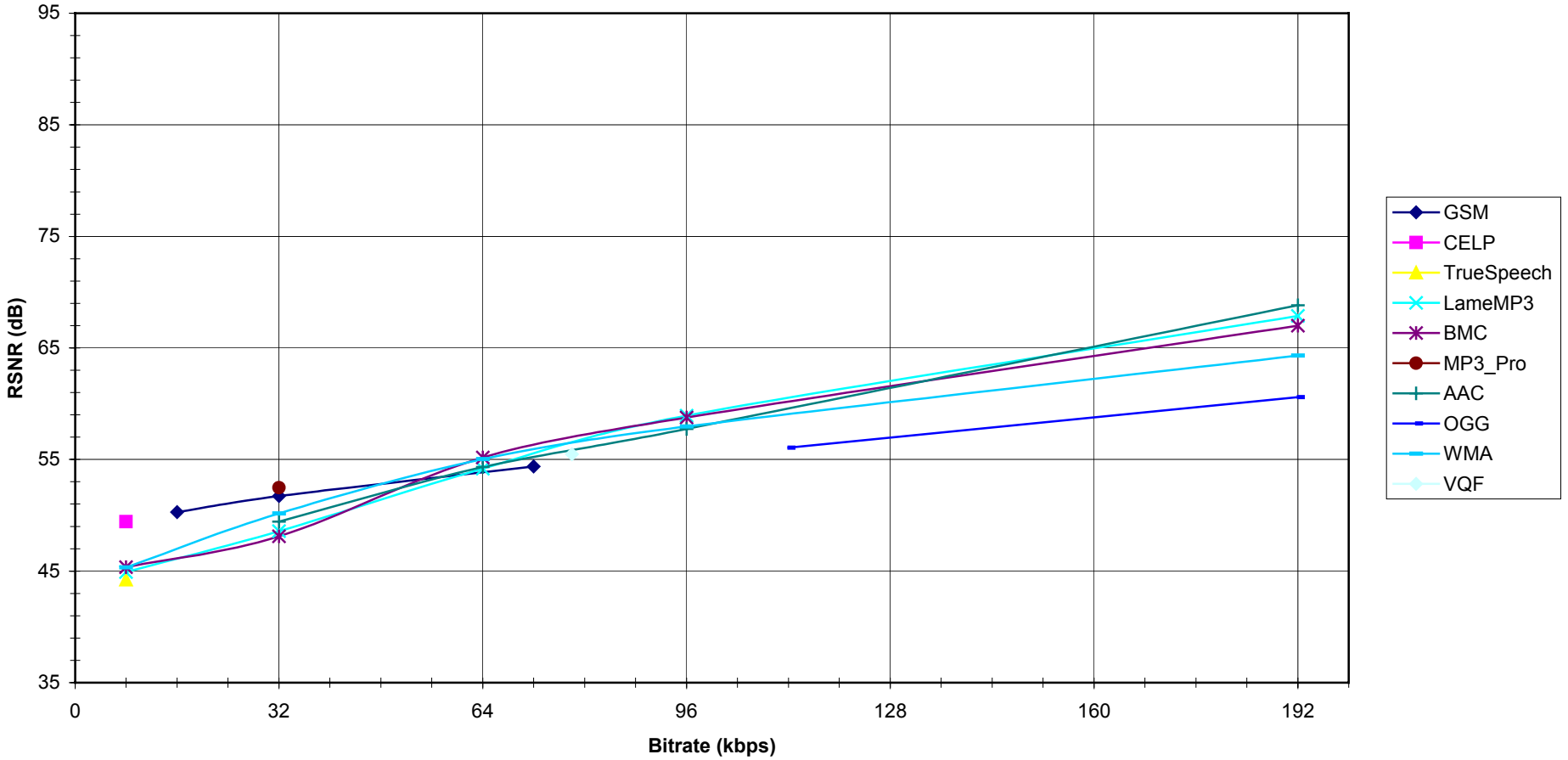


Диаграмма 4.3

PSNR средних частот файла music.wav

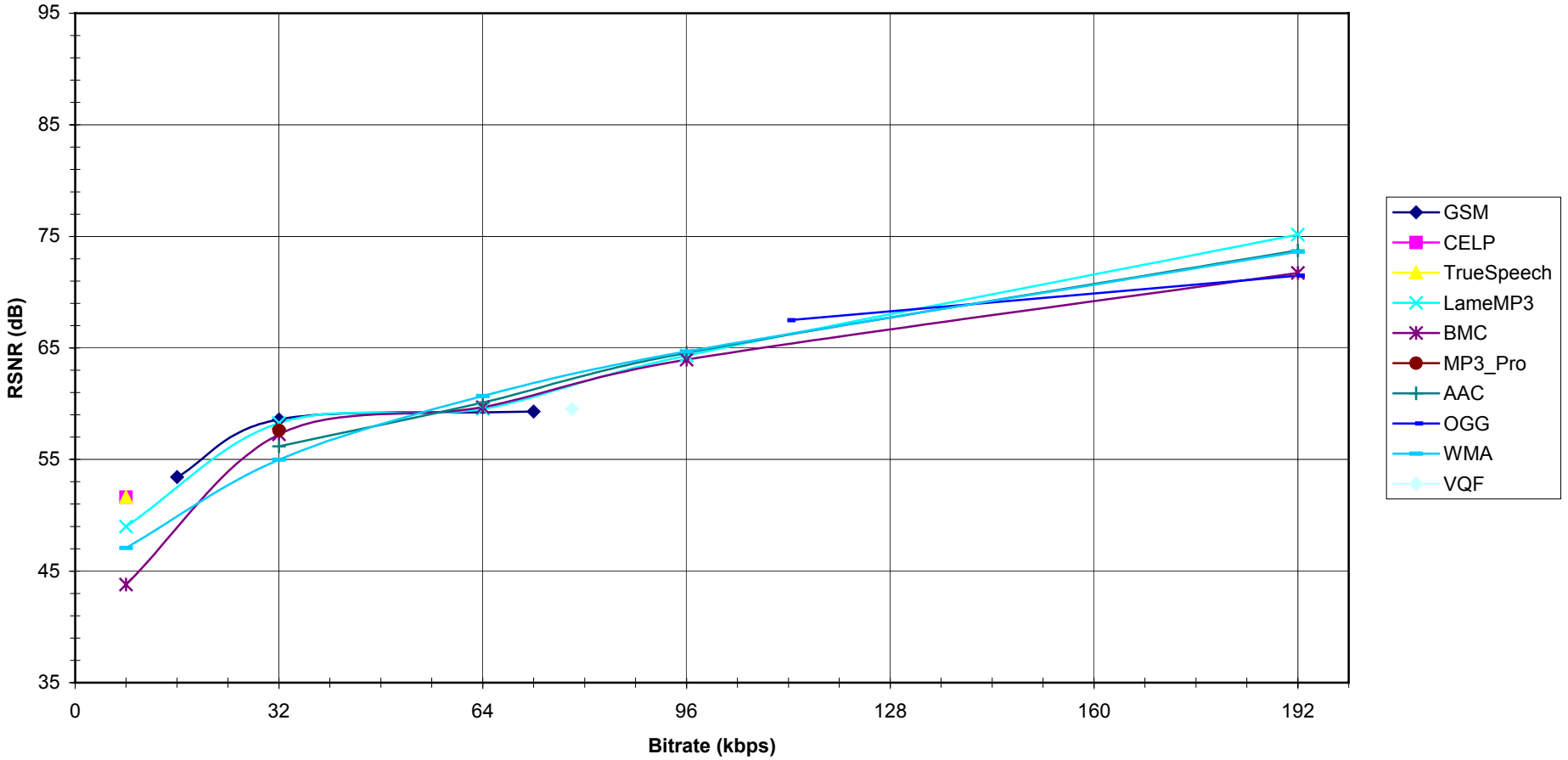


Диаграмма 4.4

PSNR низких частот файла music.wav

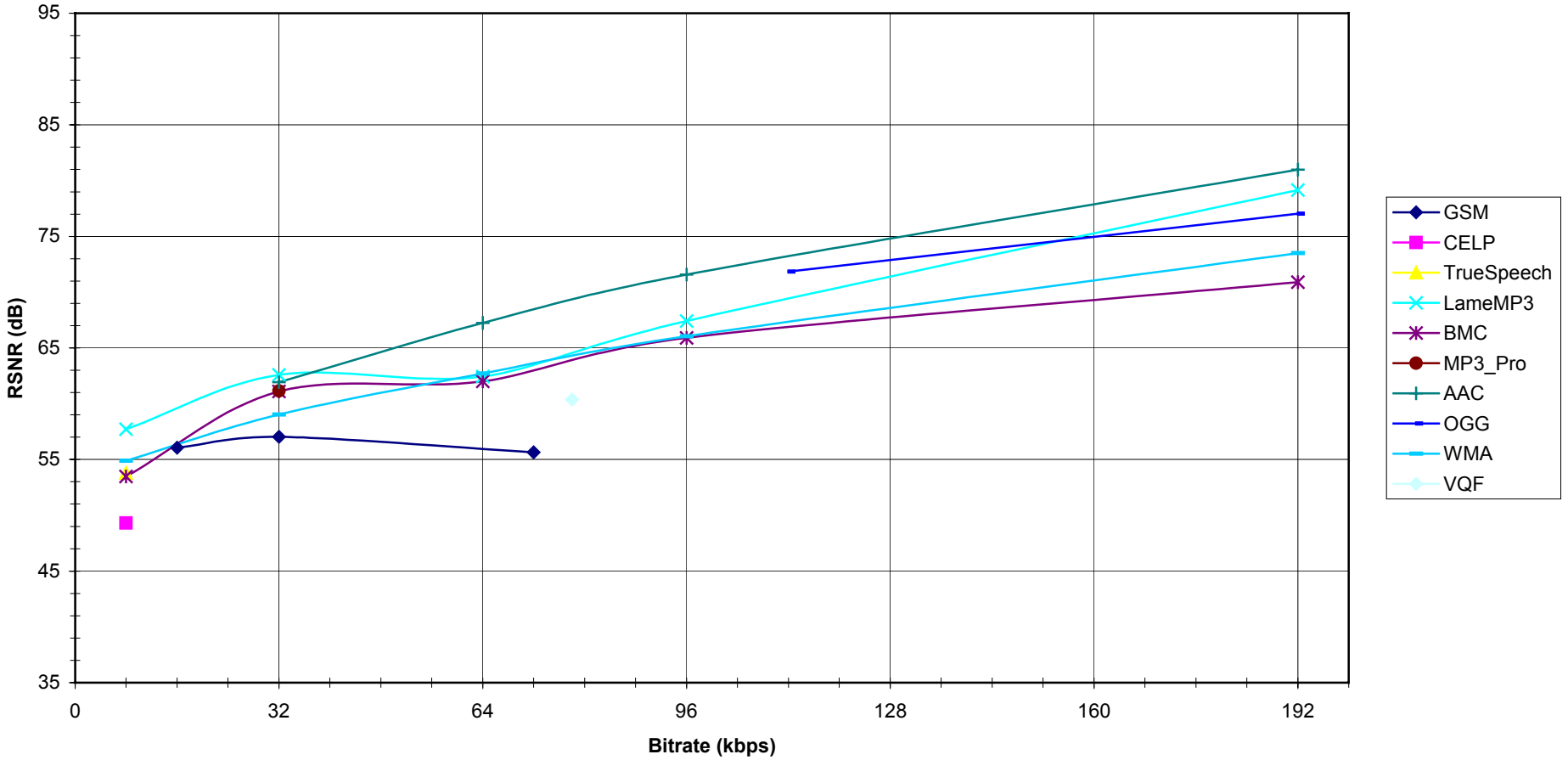


Диаграмма 5.1

PSNR файла instrvoice.wav

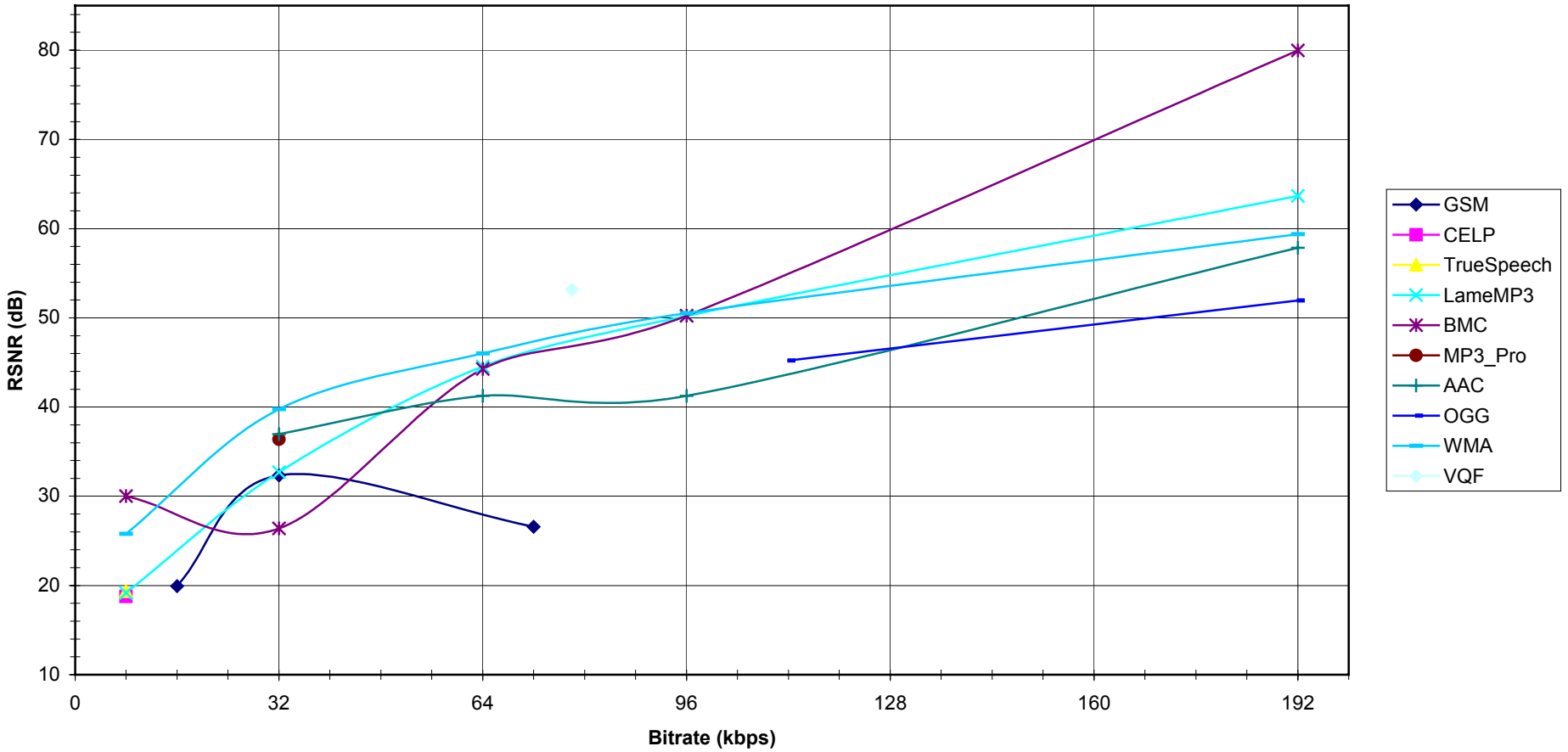


Диаграмма 5.2

PSNR высоких частот файла instrvoice.wav

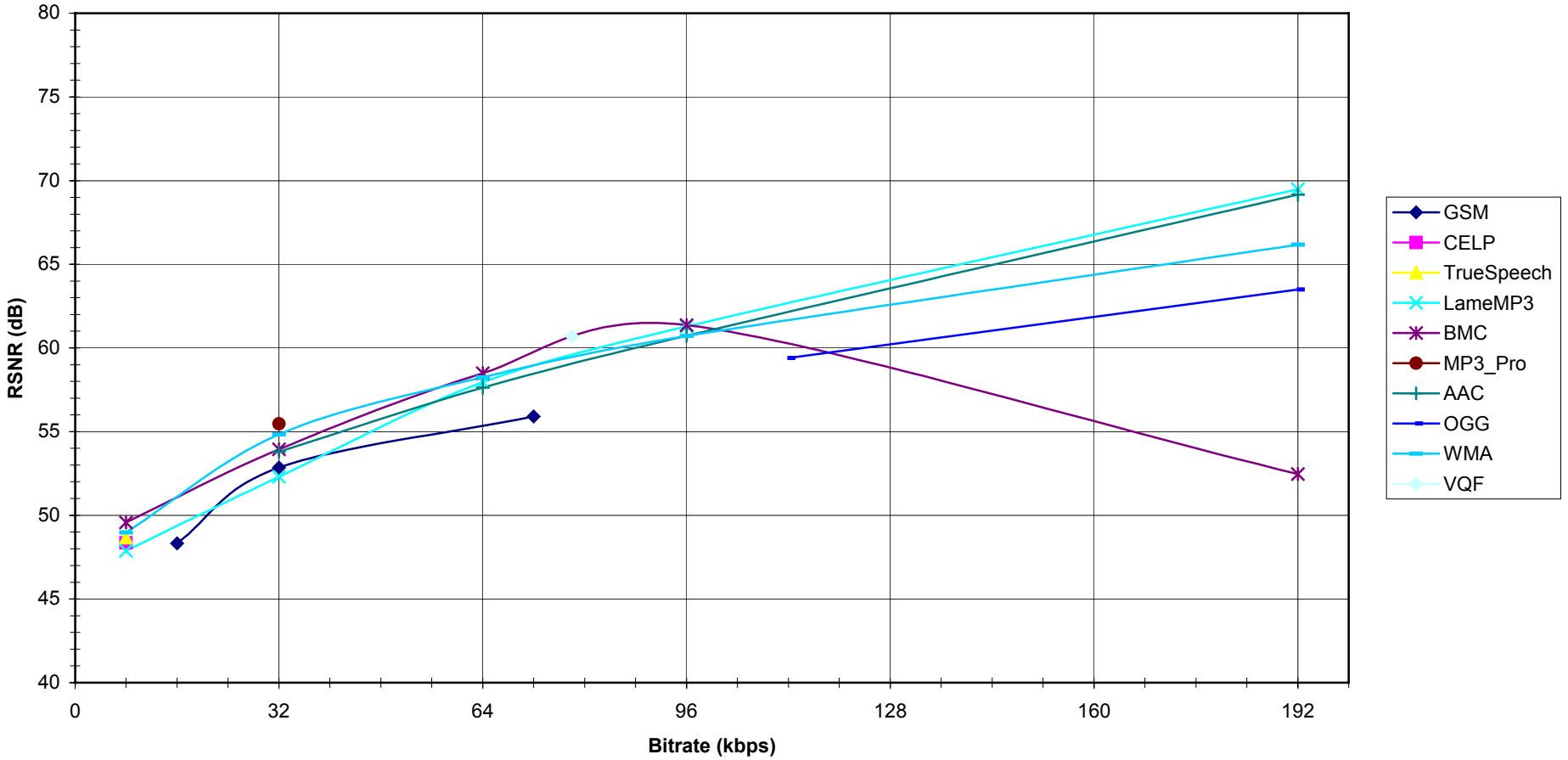


Диаграмма 5.3

PSNR средних частот файла instrvoice.wav

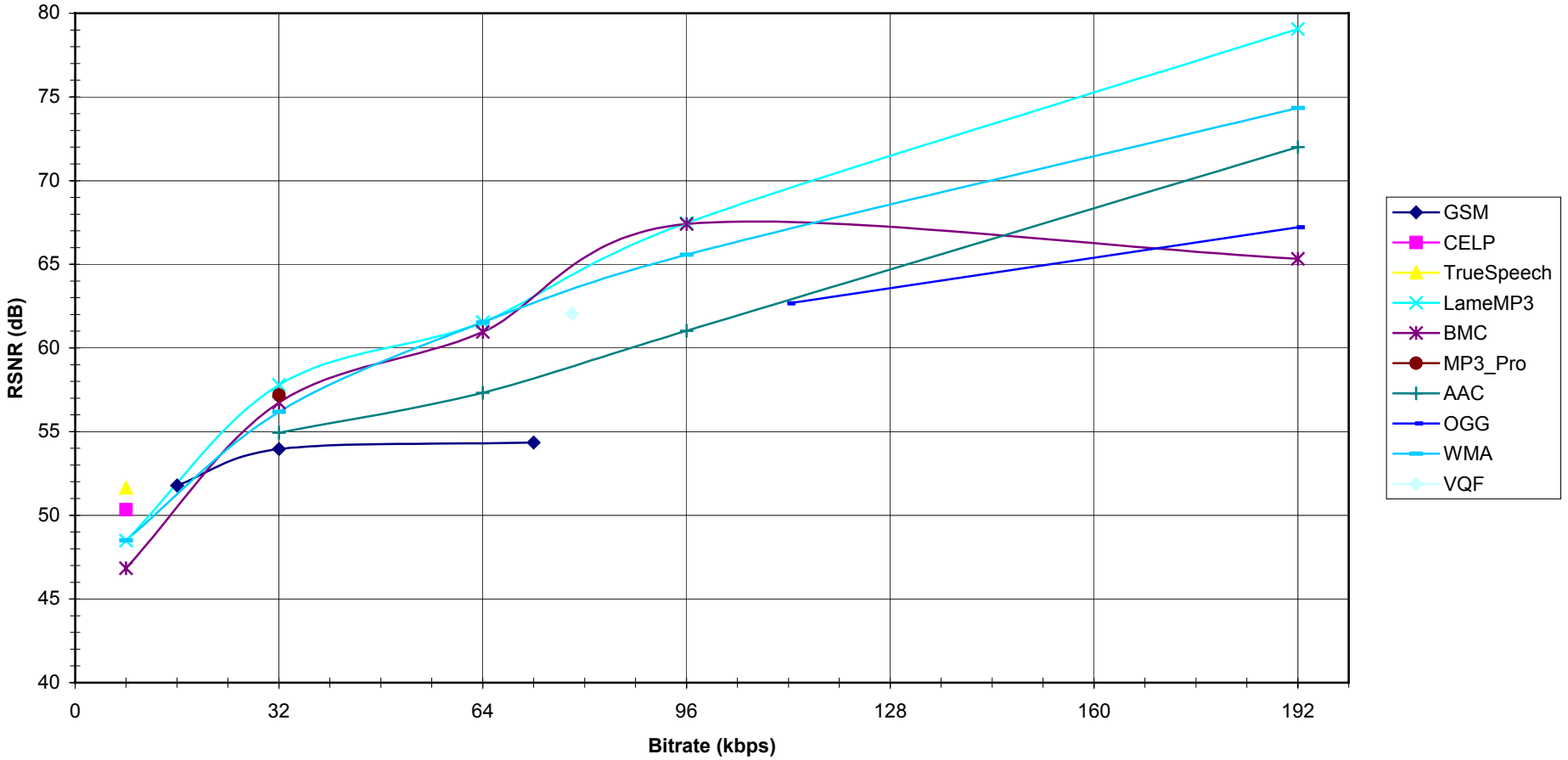
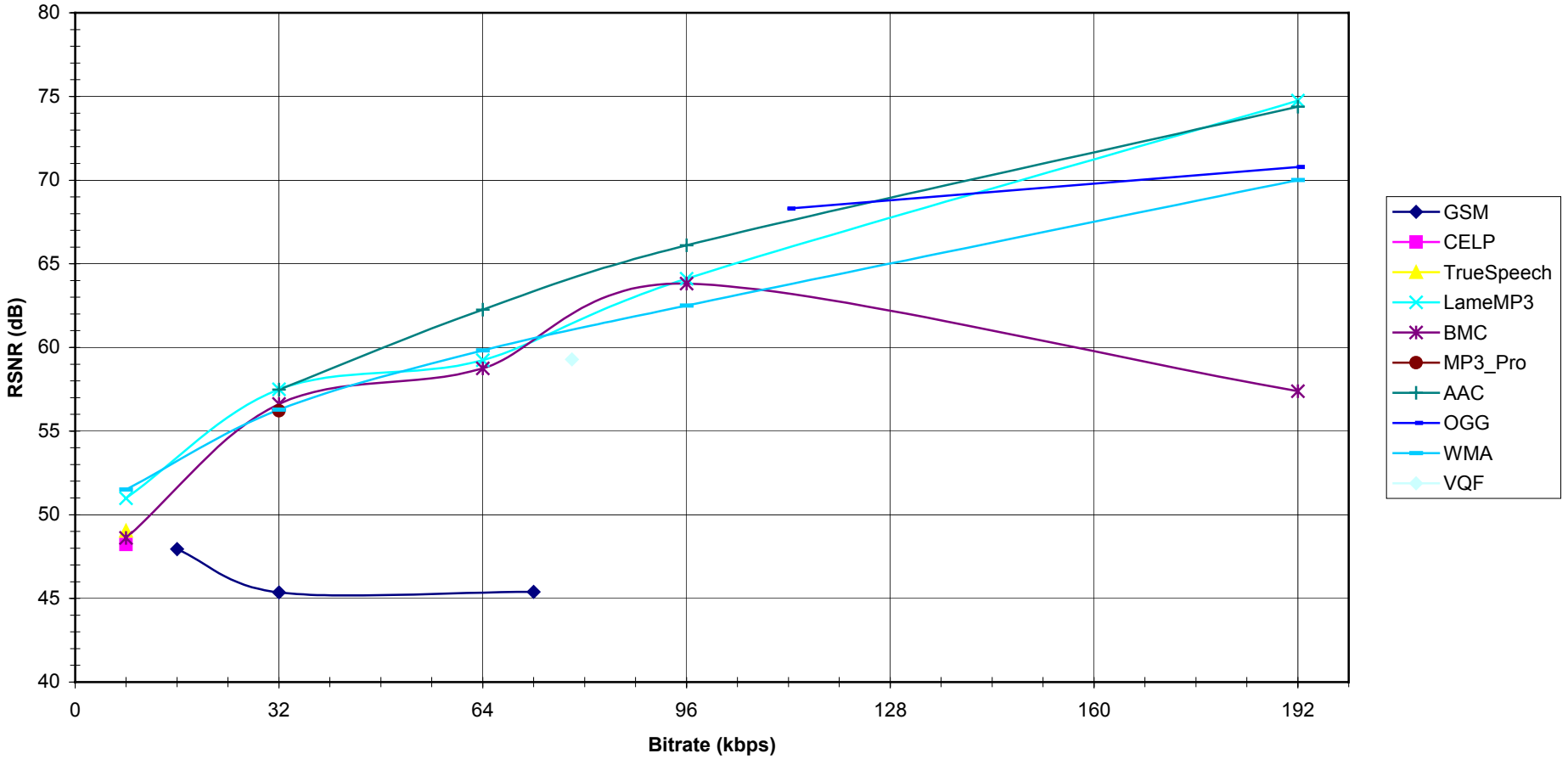


Диаграмма 5.4

PSNR низких частот файла instrvoice.wav



4. Спектральная визуализация кодеков с высоким битрейтом

4.1 Визуализация “Разницы спектров и сонограмм сигналов”

В данном разделе представлены спектры сигналов четырех тестируемых кодеков на битрейте 192 kbps для файла music.wav, построенные с использованием быстрого преобразования Фурье с окном 2048 семплов. Чем выше находится линия спектра, тем громче звук на данной частоте. Объединение синей и красной области дает спектр исходного сигнала, а синей и зеленой областей – спектр сигнала, обработанного кодеком. Данный битрейт достаточно высок, поэтому спектральные отличия сигналов оказываются малы и присутствуют, в основном, в высокочастотной области, где человек наименее чувствителен к звуку. Результаты данного теста нельзя использовать для оценки качества сохранения исходной звуковой информации напрямую, так как общие спектры композиции являются слишком усредненными данными о звуке, и мелкие, но хорошо слышимые артефакты вносимые кодеками, будут незаметны при сравнении по спектру. Тем не менее, изучение спектров позволяет понять некоторые особенности работы кодеков – такие, как порог фильтрации высоких частот на разных битрейтах, особенности сохранения общей энергии сигнала на разных частотах и т.п.

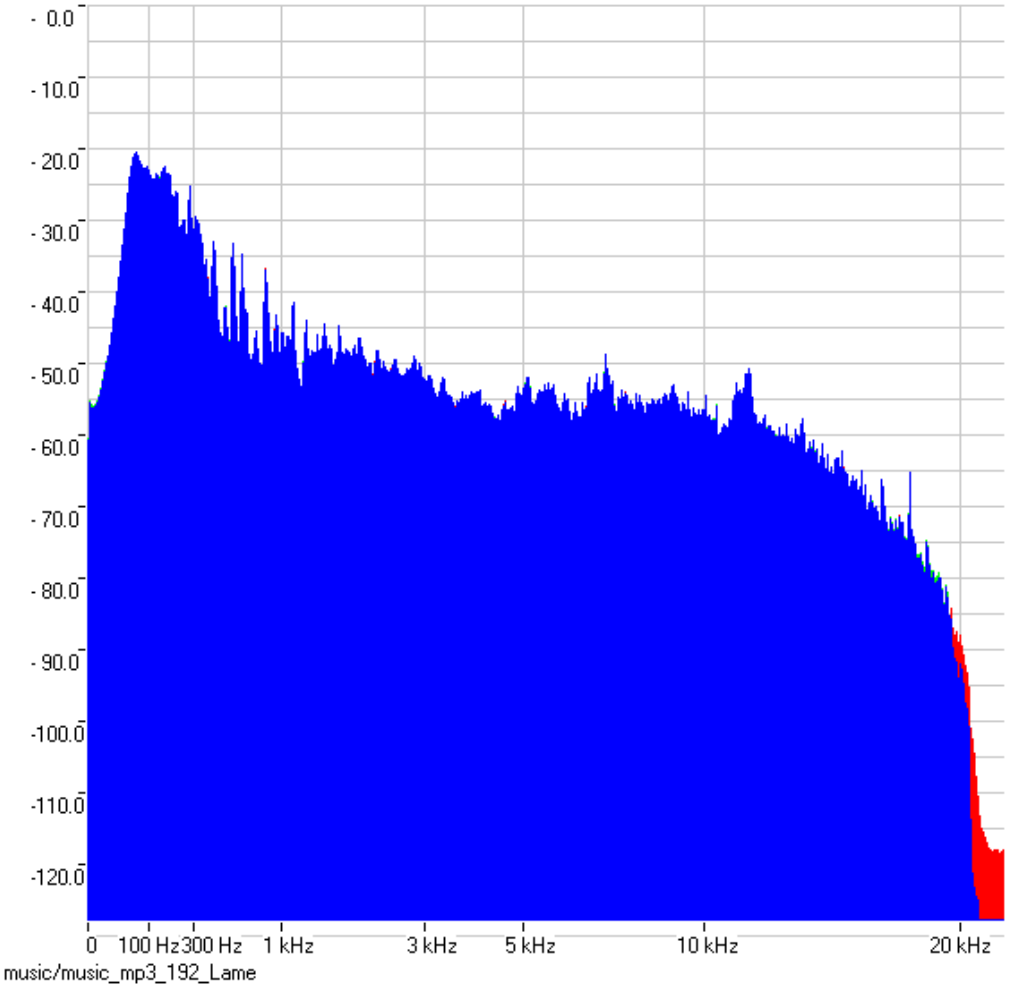


Рис 4.1.1 Lame MP3

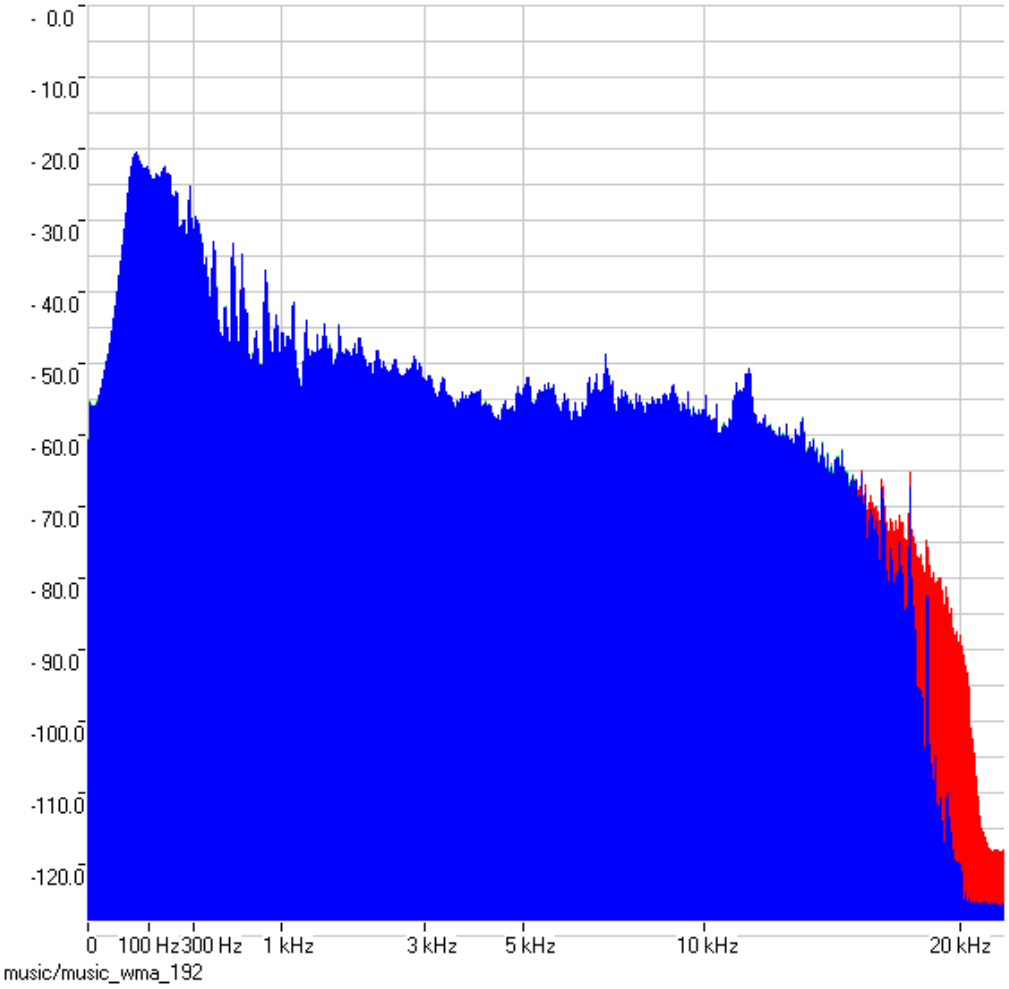


Рис 4.1.2 WMA

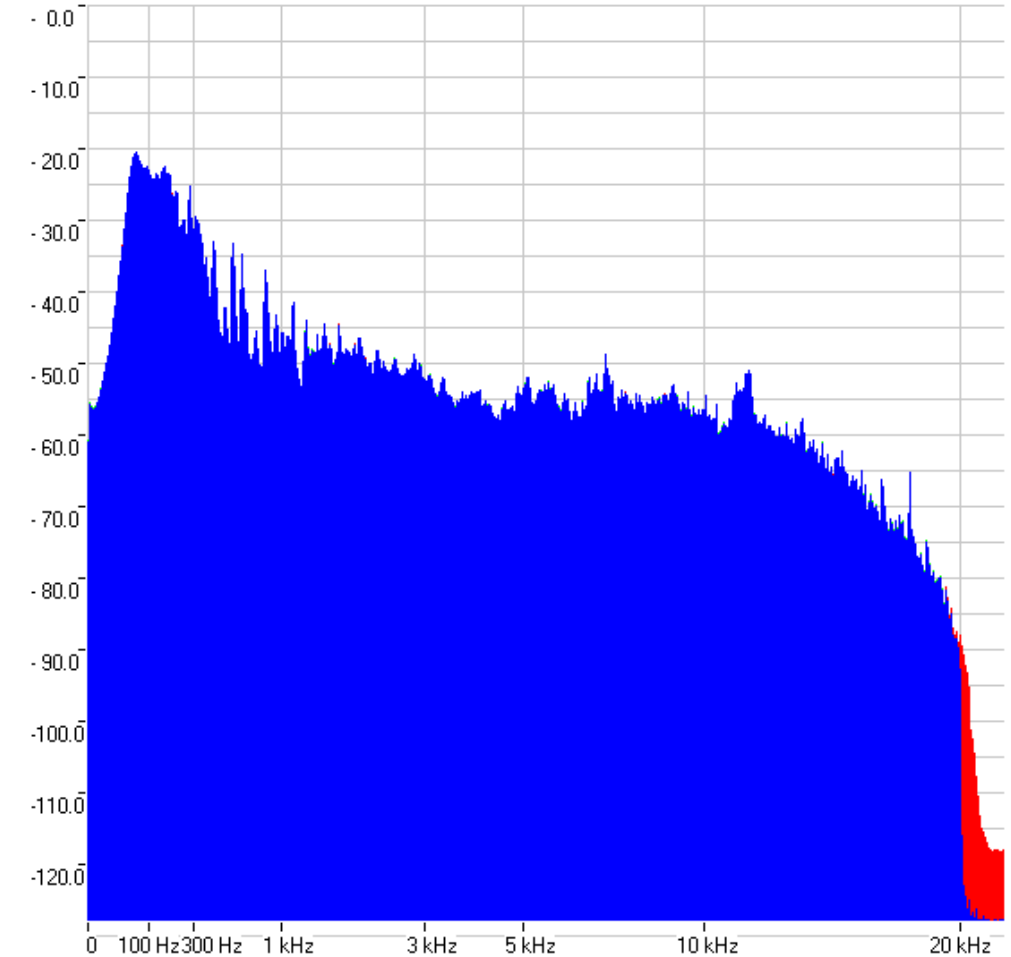


Рис 4.1.3 AAC

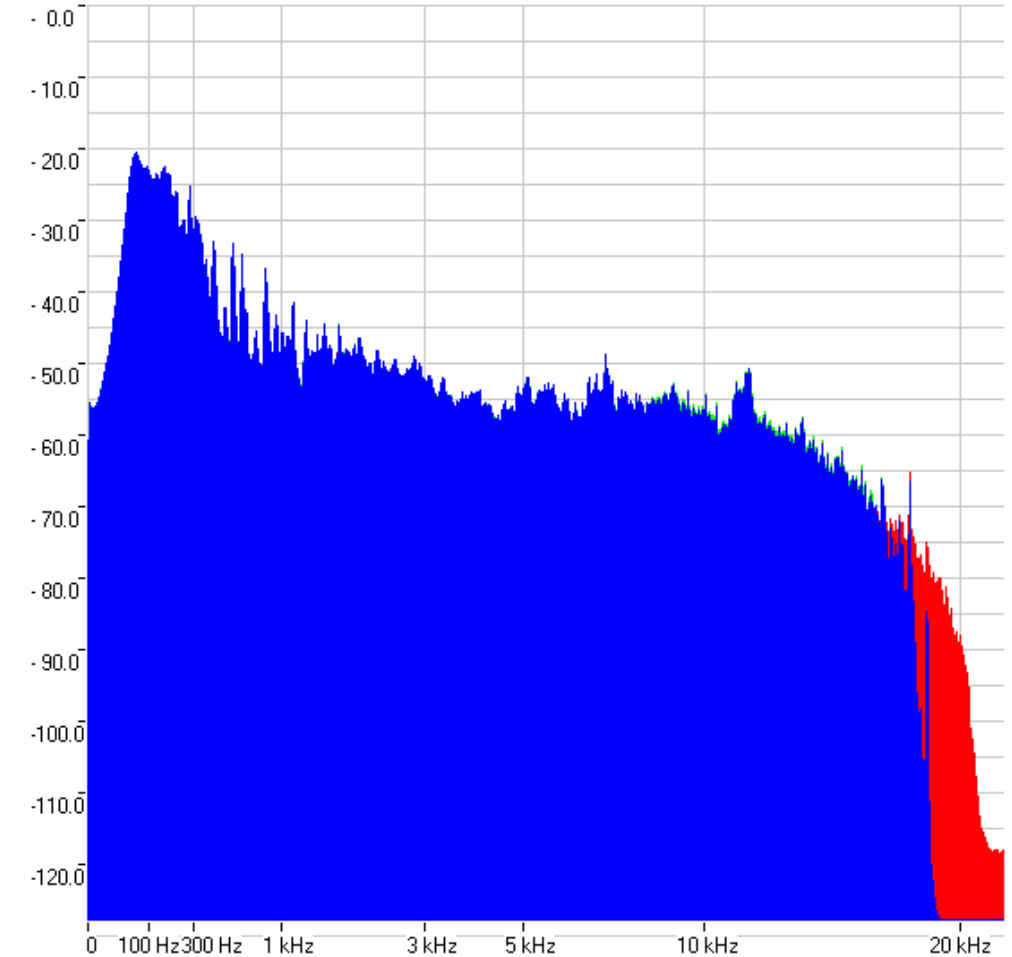


Рис 4.1.4 OGG

Видно, что на высоких битрейтах энергия верхних частот была не сохранена только в WMA и OGG, все остальные кодеки сохранили очень близкий спектр. Кодек OGG отличается тем, что некорректно сохраняет энергию в частотном диапазоне 8-17 кГц, которая увеличена по сравнению с энергией исходного сигнала (зеленые области). Поскольку тестирование по PSNR крайне чувствительно к изменению общей энергии сигнала, становится понятно, почему на диаграммах с частотно-временным PSNR для высоких частот кодек OGG показывает стабильно плохой результат. Поскольку кодек не сохранил общую энергию сигнала во всей высокочастотной области, то результаты тестирования по PSNR для данного кодека нельзя считать адекватными. Для более подробного изучения энергетических отличий в двух сигналах можно прибегнуть к визуализации трехмерных графиков, показывающих разницу сонограмм, построенных путем вычитания значений амплитуд одной сонограммы из другой в каждой точке. Такой метод позволяет увидеть некоторые дополнительные отличия в сигналах. На следующих графиках представлены: исходная сонограмма (рис 4.1.5), графики разницы сонограмм файла test.wav и результатов обработки этого файла, полученных с помощью кодеков Lame mp3, BMC mp3, AAC, OGG, WMA. Белые области на графиках разницы сонограмм свидетельствуют о том, что энергия сигналов одинакова. Синие области говорят об увеличении в данных точках мощности обработанного сигнала по сравнению с исходным, а красные об уменьшении.

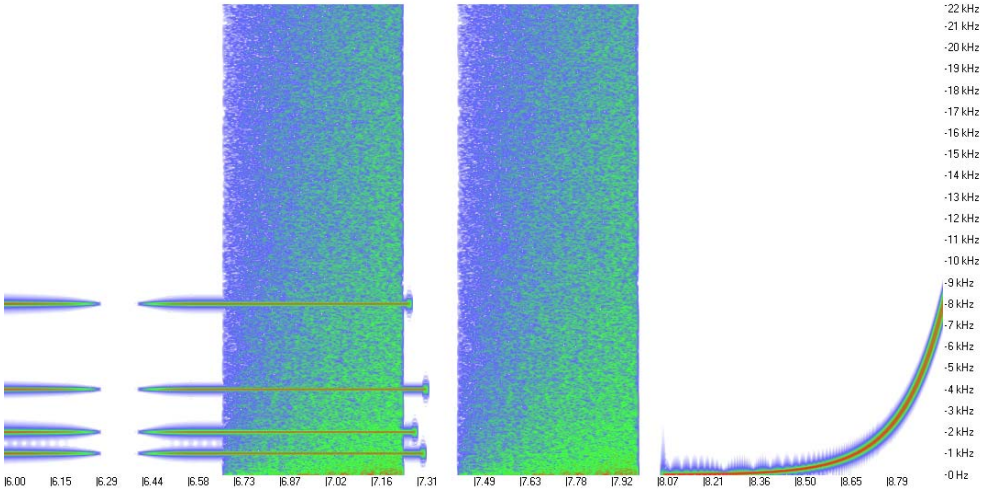


Рис 4.1.5 Исходный файл

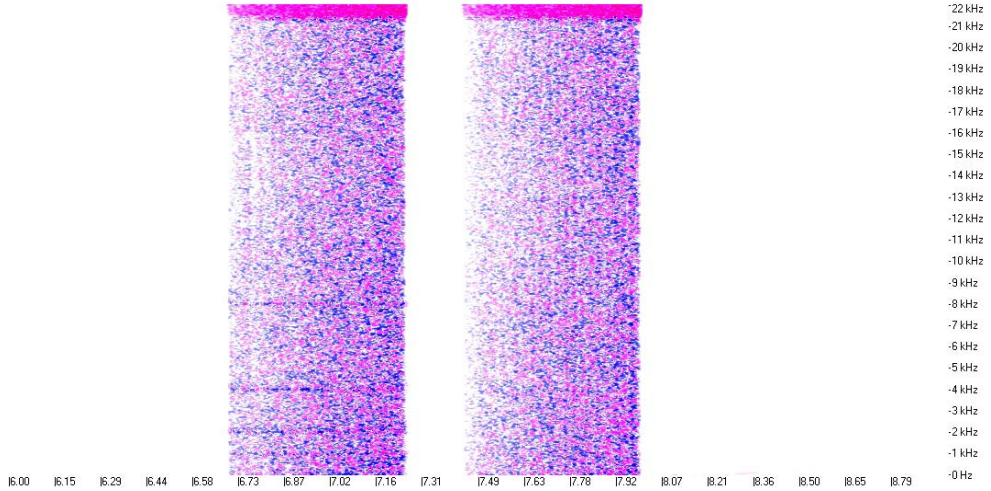


Рис 4.1.6 Lame MP3

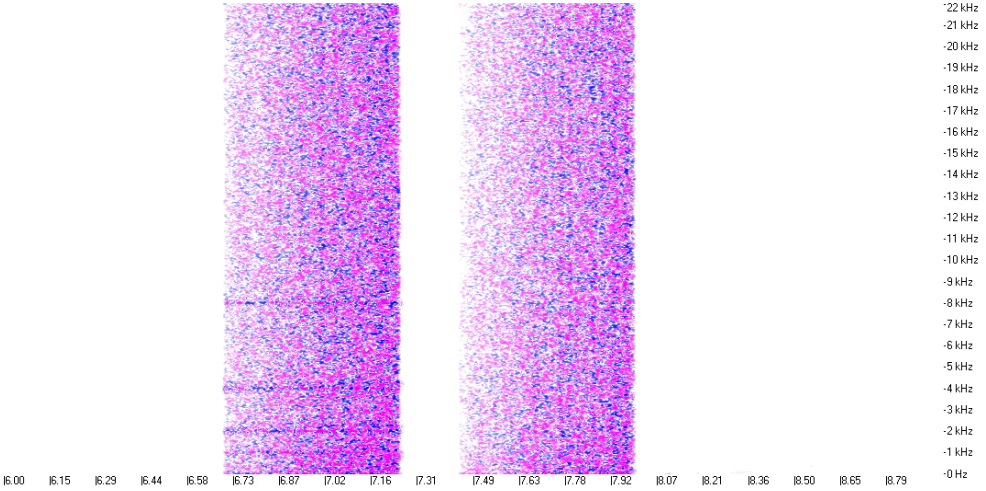


Рис 4.1.7 BMC MP3

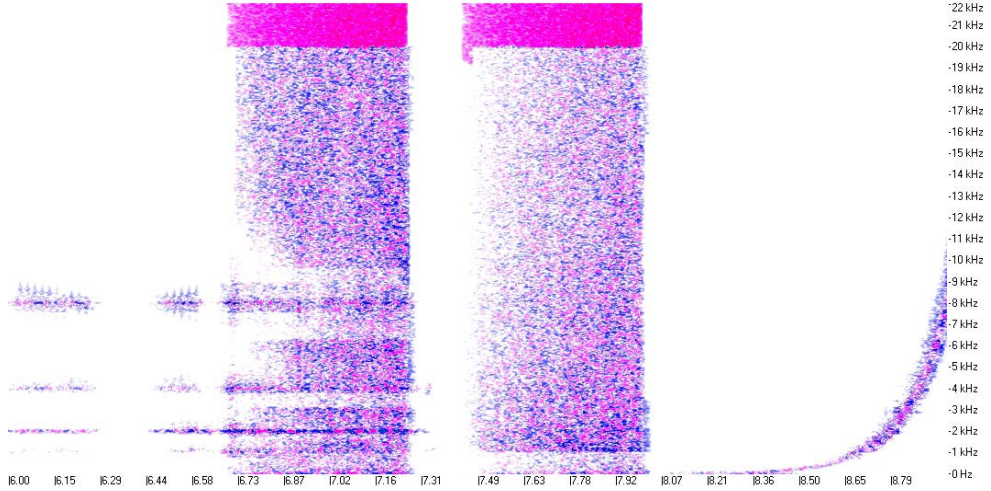


Рис 4.1.8 AAC

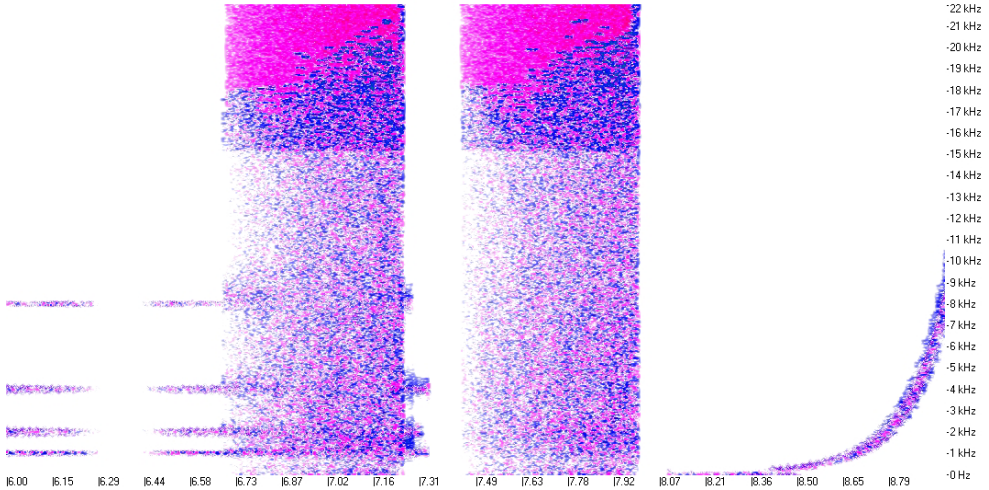


Рис 4.1.9 OGG

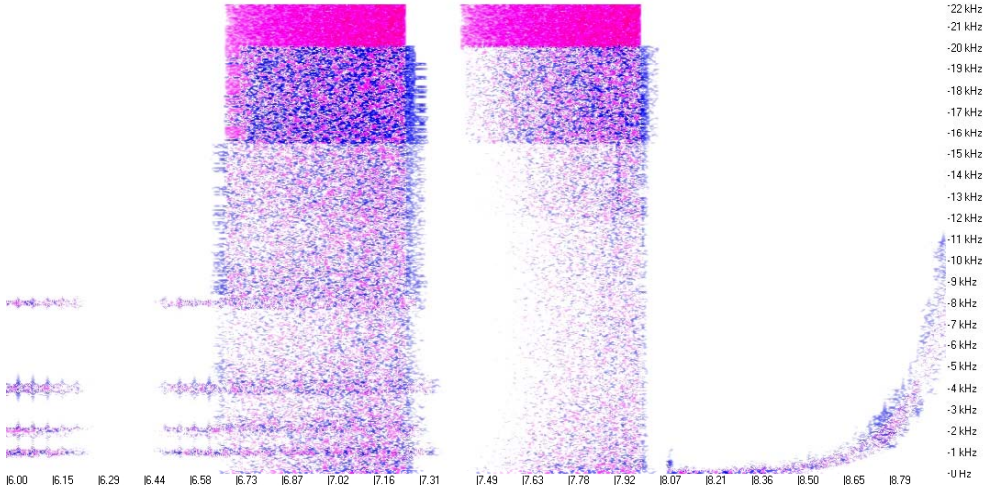


Рис 4.1.10 WMA

Тест по сонограмме проводился на файле test.wav (приведена шестая секунда).

Как видно, Lame обрезал самые высокие частоты, а в остальном диапазоне наличие сильных четких сигналов не повлияло на сонограмму разницы сигналов. Практически такая же картина и у BMC mp3, но, в отличие от Lame, высокие частоты у него не отфильтрованы. Отмечается полное сохранение энергии на громких гармонических сигналах как у Lame, так и у BMC.

Изменения сонограмм у AAC и OGG более интересны и связаны с использованием сложных психоакустических моделей.
Для AAC характерны слабые отличия в области 100-1000 кГц для спектра всего файла и не зависят от его содержимого. Видно, что наименьшие изменения проявляются в тех областях, где присутствуют четкие сигналы на фоне шума, так как при тестировании просто областей с равномерным шумом они не наблюдались.
OGG дает достаточно равномерное распределение искажений по частотам, которые увеличиваются на высоких частотах и минимальны в области 100-1000 кГц. OGG лучше всего сохраняет спектральную картину в тех местах, где после тишины появляется шумовой сигнал.

Очень интересна структура отличий сигнала, обработанного кодеком WMA, который имеет наиболее слабые искажения в области шума, и несколько более сильные – в области присутствия мощного гармонического сигнала (Сильнее отличий, которые были замерены у MP3, но значительно слабее, чем у OGG и AAC). Характерной особенностью кодека оказалось изменение сигнала на частотах выше 15 кГц, он получился немного запаздывающим относительно своей низкочастотной составляющей. Очевидно, что кодек использует различные стратегии обработки и сжатия сигнала в зависимости от частоты.

4.2 Визуализация “спектра разницы сигналов”

Для дополнительной визуализации результатов работы временной PSNR-метрики был придуман следующий алгоритм: для каждого временного отсчета из исходного сигнала вычитался сигнал, обработанный кодеком, после чего строился его спектр в логарифмическом масштабе. Данный метод визуализации позволяет найти частотные области, в которых кодек наиболее близко сохраняет подобие сигналов как в амплитудной, так и в фазовой оставляющих, что часто тоже весьма важно.
Зеленые и красные линии – спектры обработанного и исходного файлов, синяя область – спектр сигнала разницы амплитуд. Для кодеков, сохраняющих форму волны, справедливо утверждение: чем ниже находится граница синей области, тем больше подобие сигналов, т.к. сбои фазы в звучании гармонических инструментов (гитара, пианино) и голоса крайне отрицательно сказываются на качестве звука при прослушивании.

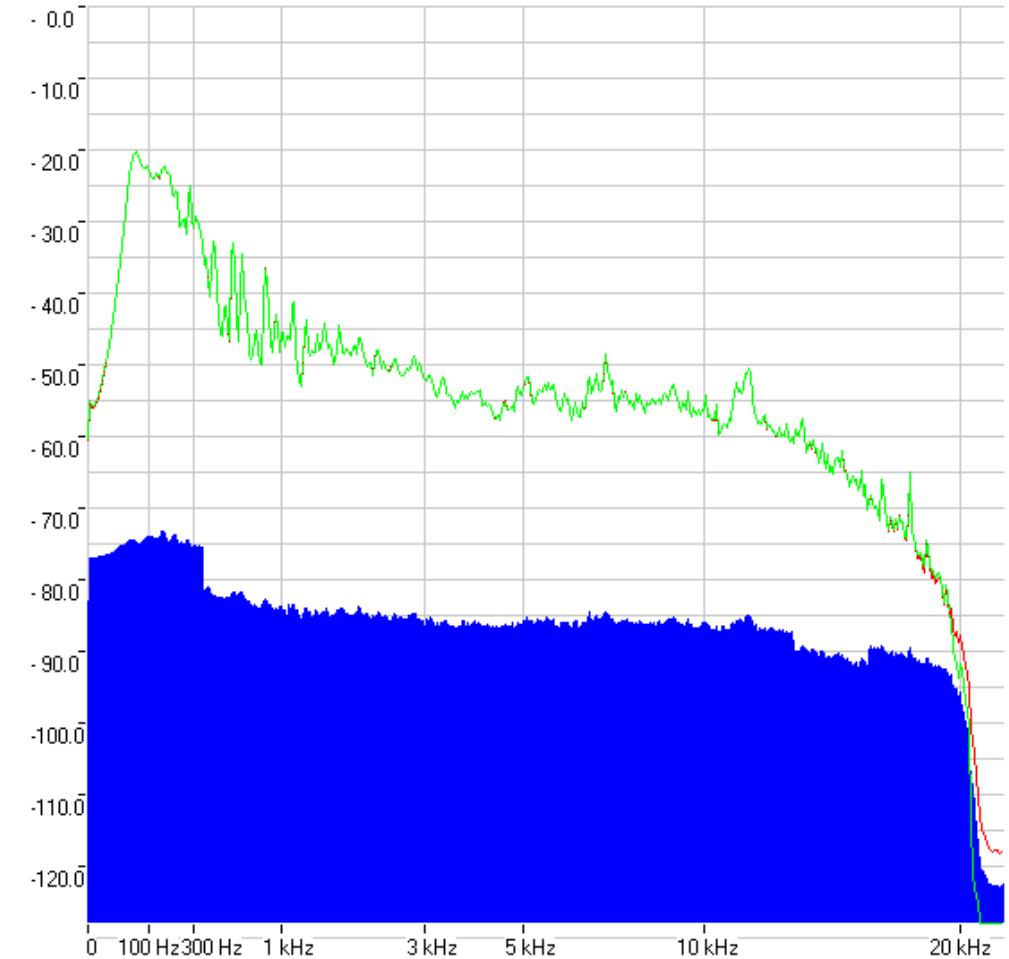


Рис 4.2.1 Lame MP3

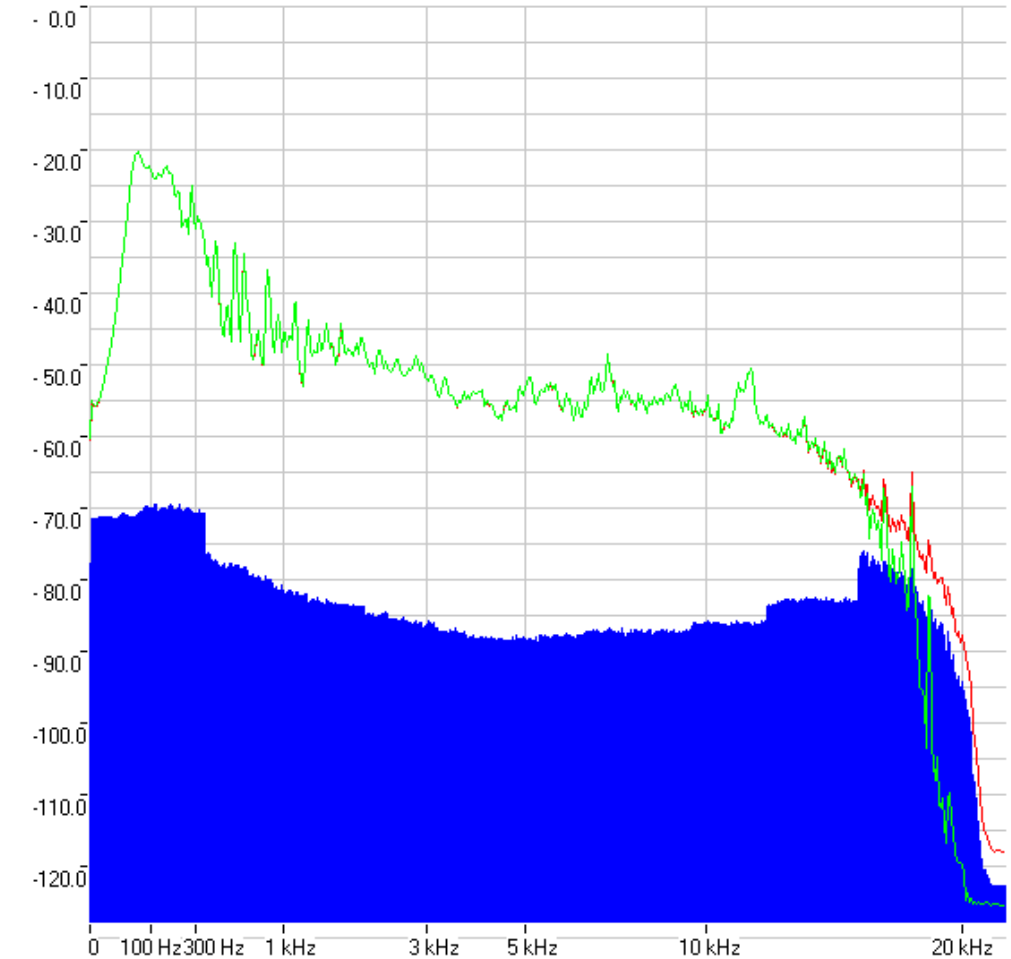


Рис 4.2.2 WMA

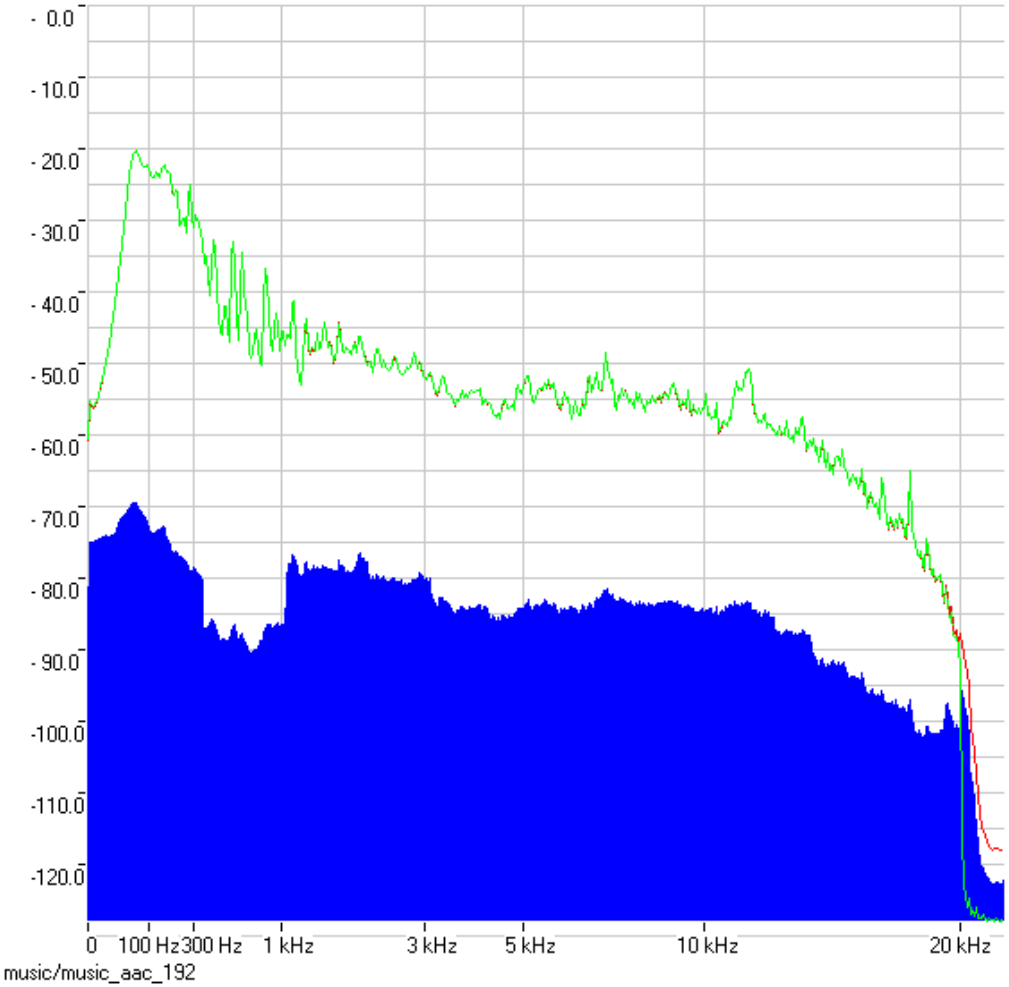


Рис 4.2.3 AAC

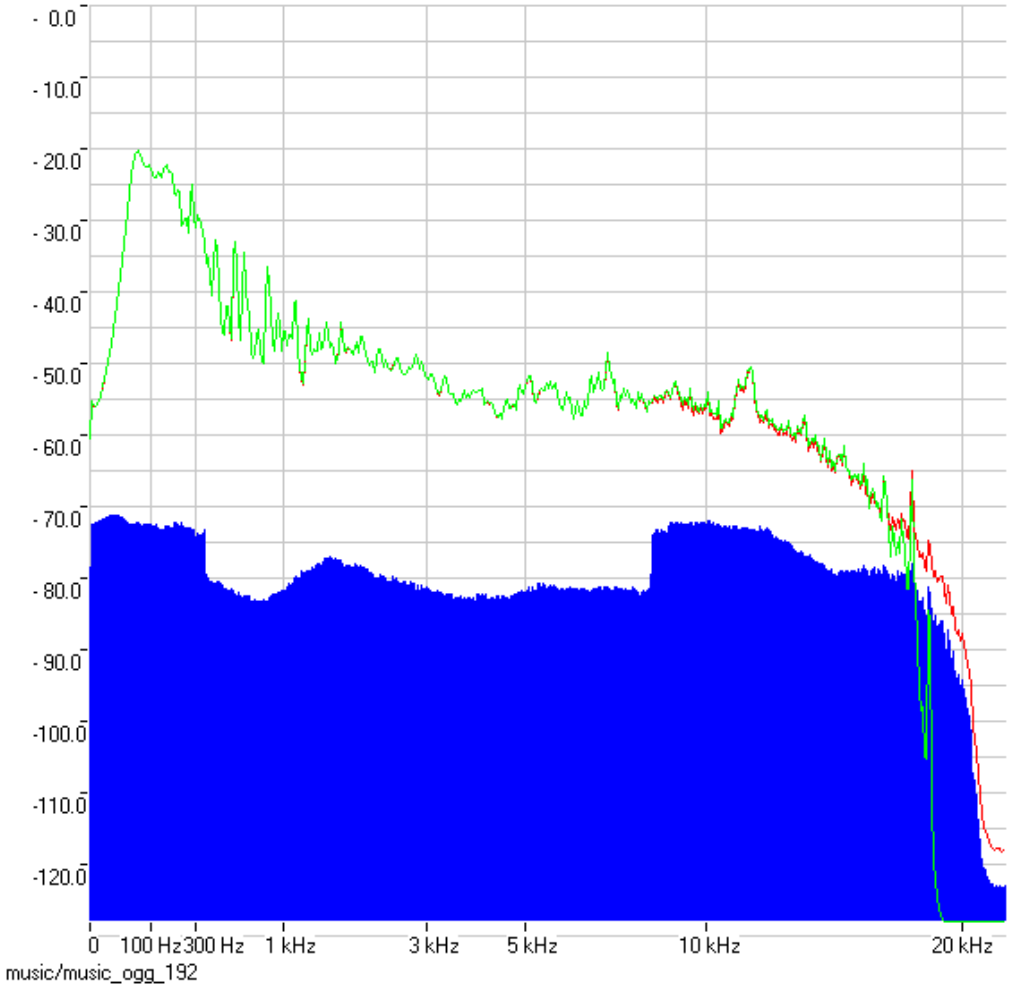


Рис 4.2.4 OGG

Такой анализ показал некоторые скрытые отличия в спектрах, которые не были до этого видны.

Спектр разницы сигнала MP3, закодированного с помощью Lame, больше напоминает спектр коричневого шума, что косвенно свидетельствует о том, что кодек примерно одинаково сохранил фазы сигнала на всех частотах. У AAC явно виден провал в области частот 100-1000Гц, что свидетельствует о том, что фазы сигнала в этом частотном диапазоне изменились меньше всего. А именно эта область наиболее различима человеческим ухом, и изменения деталей в ней хорошо заметны на слух, а также, в этой области находится основной диапазон голоса и многих гармонических музыкальных инструментов. Вероятно, такое поведение кодека обусловлено заложенной в него психоакустической моделью человеческого слуха. OGG тоже использует похожую на AAC стратегию обработки сигнала, четко виден такой же провал в области 100-1000Гц. Хорошо сохранен спектр на частотах до 8 кГц, а после восьми килогерц отклонение от оригинала сильно увеличивается, что по-прежнему связано с некачественным сохранением распределения энергии в этом диапазоне.

Кодек WMA, видимо, также использует некоторую психоакустичекую модель, но имеет несколько отличающуюся картину: у него область наиболее точного сохранения фаз сигнала лежит в районе 5 кГц, и имеются большие отличия на высоких частотах, что вполне логично, т.к. энергия в высокочастотном диапазоне не сохранена.

Анализ данных графиков позволяет утверждать, что наиболее качественно психоакустика учитывается в OGG и AAC.

5. Результаты и выводы PSNR тестирования

Основным практическим результатом исследования явилось получение графиков зависимости временного PSNR и графиков, иллюстрирующих сравнения сонограмм и спектров. Замечания, сделанные в пункте 4 данного обзора, не принимались во внимание ввиду неоднозначности их интерпретации, тем не менее, учитывая их существование, необходимо отметить, что по результатам тестирования по PSNR нельзя сделать однозначные выводы о качестве работы тех или иных звуковых кодеков.

Необходимость построения звуковой метрики

Наши эксперименты показали, что при низких битрейтах звуки достаточно искажаются и количественно оценить различные артефакты (шипение, эхо, скрежет) достаточно сложно, если вообще возможно, т.к. неясно, какой же звук хуже – «шипящий» или «булькающий». Здесь все зависит от исходного файла, поэтому «умная» метрика должна учитывать особенности кодируемого файла при сравнении его с оригиналом. С другой стороны на достаточно больших битрейтах многие люди уже перестают находить отличия в качестве кодирования, и тогда снова непонятно, как именно надо строить метрику. Поэтому наиболее корректным является использование построенных метрик для сравнения файлов со средними битрейтами - 32-128.

Корреляция между частотными и временными PSNR

Большинство графиков показывает, что между частотными и временными PSNR есть зависимость, но ее характер достаточно сложен. Временной PSNR должен был бы играть роль некоего интегрального показателя для трех частотно-временных PSNR, но результат работы метрик на файле speech.wav этого не подтверждает. У CELP и TrueSpeech одинаковые показатели по PSNR на высоких и средних частотах, на низких чуть лучше TrueSpeech, но по временному PSNR намного лучше CELP. Вероятнее всего это связано с принципиальными отличиями в стратегиях сохранения формы волны, которые сказываются на результатах тестирования по временному PSNR.

Обобщенные выводы тестирования, однозначно определяющие кодек – лидер, получить невозможно, т.к. результат изменяется как в зависимости от типа звука во входной последовательности, так и выбора частотного диапазона, где необходимо, чтобы кодек имел наилучшие показатели.

Для определенности приведена таблица лидеров тестирования только на файле Test.wav, т.к. в нем подробно известна структура звука, вплоть до формы волны, и результаты тестирования по PSNR для различных кодеков имеют наибольшую разницу, сохраняя стабильность показаний на разных битрейтах.

Таблица 5.1 (Кодеки-лидеры по результатам PSNR тестирования для файла Test.wav)

Файл/график	Низкие битрейты	Высокие битрейты
	8-64	112-192
Test.wav		
PSNR (диагр.1.1)	WMA	BMC
ВЧ (диагр.1.2)	WMA	BMC
СЧ (диагр.1.3)	Lame	Lame
НЧ (диагр.1.4)	WMA	Lame

Показатели частотно-временных PSNR для средних и высоких частот действительно выше у речевых кодеков, но, как уже упоминалось, на качество звука это должным образом не повлияло. На низких частотах речевые кодеки немного уступают лидеру временного PSNR – WMA, хотя на слух 8-бит WMA очень похож на MP3.

Особенности кодеков

Для Lame MP3 характерна линейная зависимость PSNR от битрейтах от 64 kbps до 192 kbps. Для меньших битрейтов линейность нарушается, и хотя на этом отрезке данных недостаточно, можно предположить, что зависимость при меньших битрейтах тоже линейна, но она имеет другой угловой коэффициент или сдвинута вниз. На большинстве диаграмм, как временных, так и частотных PSNR, график Lame MP3 претерпевает «изгиб» на отрезке 32-64 kbps. Аналогичным образом выглядят результаты для BMC MP3. Это позволяет предположить, что данная особенность характерна для всего стандарта mp3.

Оправданность использования речевых кодеков

Полученные данные не позволяют нам утверждать, что специализированные речевые кодеки действительно необходимы для сохранения речи. Ни при одной из реализованных метрик их преимущество не было очевидным, в том числе и после внимательного прослушивания закодированного файла speech.wav и сравнения с оригиналом.

При прослушивании восьми-килобитных файлов грубые искажения голоса были слышны у всех кодеков, но проявлялись они по-разному. У вокодеров (скорее всего из-за не самого удачного расширения частотного диапазона) звук получался полным скрипа и скрежета. После MP3 файл звучал своеобразно глухо – как будто человек говорил в вату. Любопытное наблюдение: если из сжатого и после разжатого файла CELP убрать все частоты выше 4-5 kHz, то неприятные скрипы пропадают, и звук получается более качественным, чем с высокими частотами. Начиная с битрейтов в 16 – 32 кбита явные искажения звука исчезают как в GSM, так и в универсальных кодеках.

Речевые кодеки, по сравнению с универсальными, обладают малой вычислительной сложностью и малым выходным потоком данных. Основные задачи, которые должны решать речевые кодеки – это сжатие речевой информации на оборудовании с малой вычислительной мощностью и последующая передача по цифровым каналам с малой пропускной способностью, что может использоваться в сотовых телефонах. С одной сторо-

ны, на сегодняшний момент производительность цифровой техники позволяет легко вычислять сложные математические преобразования, ввиду чего возможна реализация кодека с более сложными алгоритмами, подобными тем, что используются в универсальных кодеках. С другой стороны, проблема передачи большого потока данных по каналам остается и сегодня, в результате чего возможной областью применения речевых кодеков остается сжатие с очень низкими битрейтами (меньше 8kbps). Универсальные кодеки практически не поддерживают сжатие со столь низким потоком данных, т.к. после их обработки останется только низкочастотный сигнал.

6 Благодарности

Авторы выражают особую признательность за помощь в проверке и подготовке данной статьи Сергею Лукьянову, Виталию Иванову, Алексею Москвину, Олегу Петрову, Артему Титаренко.