СРАВНЕНИЕ ПОДХОДОВ К ПОСТРОЕНИЮ ХРОМАГРАММЫ ДЛЯ АНАЛИЗА АУДИО

Каминский А. В., Петровский Н. А.
Кафедра электронных вычислительных машин,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: sasha.pinsk2003@gmail.com, nick.petrovsky@bsuir.by

В работе проведён сравнительный анализ хромаграмм, построенных на основе мел-спектрограммы и константа-Q преобразования (CQT). Показано, что мел-хромаграмма плохо выделяет отдельные музыкальные ноты: энергия слишком равномерно распределена, границы нот не различимы. В отличие от неё, CQT-хромаграмма обеспечивает чёткое различие нот и хорошо отражает последовательность мелодии. Результаты оценивались визуально, демонстрируя преимущество CQT для анализа музыкальных сигналов с отдельными нотами.

Введение

Музыкальный сигнал представляет собой сложный нестационарный процесс, содержащий одновременно частотную, временную и гармоническую информацию.

Для анализа музыкального контента — таких задач, как определение аккордов, тональности, а также систем поиска по напеву (Query-by-Humming) — требуется устойчивое и компактное представление аудио, отражающее свойства звука независимо от тембра и октавы. Одним из наиболее распространённых способов такого представления является хромаграмма [1] (англ. chroma feature), описывающая распределение энергии по двенадцати высотным компонентам (C, C#, D, . . . , B) в каждый момент времени.

Построение хромаграммы основывается на частотно-временном анализе сигнала. Наиболее известными подходами являются использование мел-спектрограммы и константа-Q преобразование (CQT, Constant-Q Transform).

Цель работы — сравнение мел- и CQT-хромаграмм при анализе музыкальных сигналов. Сравнение выполнено теоретически и экспериментально на одном аудиофайле.

I. Теоретические основы

Мел-спектрограмма [2] представляет собой преобразование аудиосигнала, полученное на основе оконного преобразования Фурье (STFT) с последующим применением банка мел-фильтров, который аппроксимирует восприятие частоты человеческим слухом.

В отличие от линейного частотного разбиения в обычной спектрограмме, мел-шкала нелинейна: она имеет высокое разрешение на низких частотах и сниженное — на высоких, что отражает особенности слухового восприятия.

Сигнал $x[n]$ делится на перекрывающиеся участки длиной N. Каждый участок умножается на оконную функцию $w[n]$ (например, Ханна или Хэмминга), после чего вычисляется дискретное преобразование Фурье.

Чтобы перейти от линейной частотной шкалы к той, которая отражает, как человек действительно воспринимает высоту звука, используют мел-шкалу:

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right).$$

На основе этой зависимости строится банк мел-фильтров [3], состоящий из треугольных фильтров $H_i(k)$, равномерно распределённых по мел-шкале.

Каждый фильтр пропускает энергию только в своей частотной области и затухает к нулю на границах:

$$E_i(m) = \sum_{k=0}^{N/2} |X(m,k)|^2 \, H_i(k), \quad i = 1,2,\ldots,M.$$

Константа-Q преобразование [4] использует окна с одинаковым отношением частоты к полосе, обеспечивая логарифмическое распределение частот, соответствующее музыкальной шкале в отличие от линейной сетки STFT.

Частоты дискретных бинов f_k определяются как:

$$f_k = f_{\min} \cdot 2^{k/B}, \quad k = 0,1,\ldots,K-1.$$

Так как частоты f_k расположены логарифмически и совпадают с музыкальными нотами, CQT естественным образом подходит для построения хромаграммы. Каждый бин k сопоставляется высотной компоненте по формуле:

$$p = \text{round}\left(12 \log_2 \frac{f_k}{f_{\text{ref}}}\right) \bmod 12,$$

где $f_{\text{ref}} = 440$ – опорная частота (нота A4).

Энергии всех бинов, относящихся к одной нотной компоненте, суммируются, в результате чего получается 12-мерный вектор — хромавектор для текущего временного кадра.

II. Реализация построения хромаграмм

В качестве исходных данных для построения хромаграммы использовался аудиофайл, содержащий вокальную последовательность, в которой

исполняются ноты в диапазоне от C4 до B4, а затем в обратном направлении.

Хромаграмма на основе мел-спектрограммы строилась в несколько этапов. Сначала аудиосигнал нормализовался по амплитуде, далее разбивался на перекрывающиеся окна фиксированной длины с применением оконной функции (Хэмминга), после чего для каждого окна вычислялось дискретное преобразование Фурье (ДПФ) с перекрытием 50%. Амплитудный спектр каждого окна преобразовывался в энергетическую форму и подавался на вход банка мел-фильтров. Результатом этого этапа являлась мел-спектрограмма, отражающая распределение энергии по восприятию человека.

В итоге для каждого временного окна формировался вектор из 12 компонент, отражающий относительную энергию нотных компонент. Совокупность таких векторов во времени образует хромаграмму, которая является компактным представлением тональной структуры музыкального сигнала.

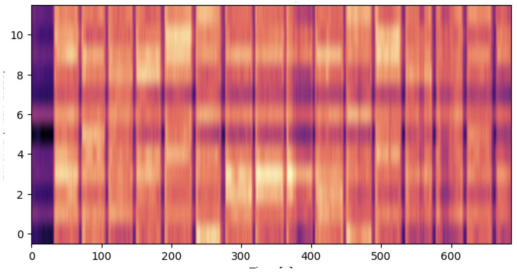

Рис. 1 – Полученная хромаграмма

Для построения хромаграммы на основе константа-Q преобразования аудиосигнал предварительно нормализовывался и разделялся на перекрывающиеся окна. Для каждого окна вычислялось константа-Q преобразование, результатом которого является спектр с логарифмическим распределением частотных бинов. Из полученной спектрограммы извлекались значения амплитуд для фиксированного числа бинов, покрывающих несколько октав. Далее для каждого бина определялась принадлежность к одному из 12 компонент высоты, соответствующих полутонам в октаве. Таким образом формировалась CQT-хромаграмма, отображающая распределение энергии по музыкальным компонентам высоты звука.

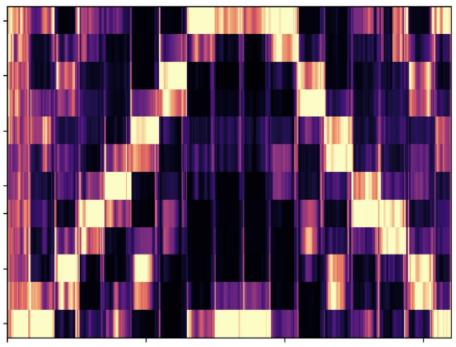

Рис. 2 – Хромаграмма после CQT

III. Результаты

Были построены хромаграммы на основе мел-спектрограммы и константа-Q преобразования. Мел-хромаграмма оказалась сглаженной и слабо различала ноты, тогда как CQT обеспечила чёткое разделение полутонов и лучшее соответствие последовательности нот.

Для количественного сравнения использовались косинусное расстояние и динамическое выравнивание по времени. CQT-хромаграмма показала большее сходство с эталоном: косинусное расстояние — 0.43 против 0.26 для мел, нормализованное DTW [5] — 0.385 против 0.735 соответственно. Это подтверждает преимущество CQT при анализе музыкальных сигналов.

1. Islam R., Tarique M. A novel convolutional neural network based dysphonic voice detection algorithm using chromagram [Electronic resource] / R. Islam, M. Tarique – University of Windsor, 2023. – Mode of access: `https://surl.li/ektwja`. – Date of access: 24.10.2025.
2. Banuroopa, K., Shanmuga Priyaa, D. A novel voiceprint using ensembled Mel-Chromagram for speaker recognition [Electronic resource] / K. Banuroopa, D. Shanmuga Priyaa. – 2022. – International Journal of Health Sciences. – Vol. 6, S4. – P. 8043–8056. – Mode of access: `https://surl.li/ukkffi`. – Date of access: 24.10.2025.
3. Vlaj, D., Kotnik, B., Horvat, B., Kačič, Z. A Computationally Efficient Mel-Filter Bank VAD Algorithm for Distributed Speech Recognition Systems / D. Vlaj, B. Kotnik, B. Horvat, Z. Kačič // EURASIP J. Advances in Signal Processing. – Vol. 2005, Article 561951 (2005). – Mode of access: `https://doi.org/10.1155/ASP.2005.487`. – Date of access: 24.10.2025.
4. Wolf-Monheim, F. Spectral and Rhythm Features for Audio Classification with Deep Convolutional Neural Networks [Electronic resource] / F. Wolf-Monheim. – 2024. – arXiv:2410.06927. – Mode of access: `https://arxiv.org/abs/2410.06927`. – Date of access: 24.10.2025.
5. Senin P. Dynamic Time Warping Algorithm Review [Electronic resource] / P. Senin – University of Hawaii at Manoa, December 2008. – Mode of access: `https://clck.ru/3Q4ohq`. – Date of access: 24.10.2025.