Музыкальная транскрипция при помощи методов машинного обучения

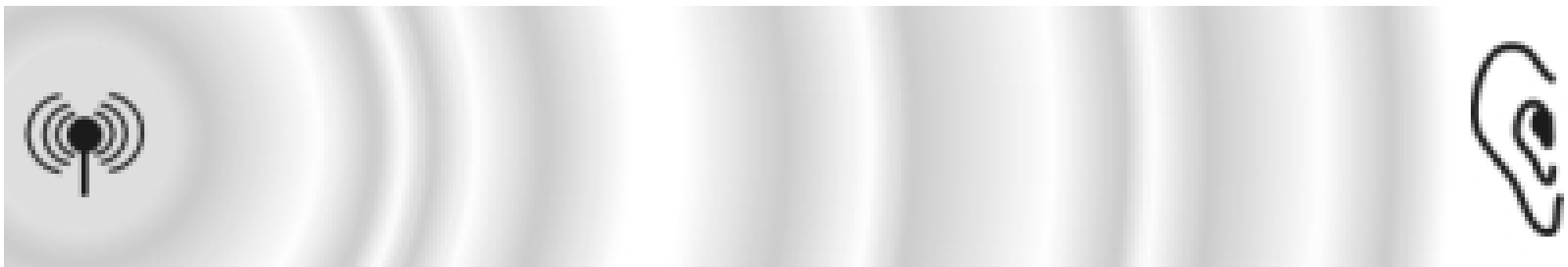
Евгений Нижибицкий

почти ВМК МГУ

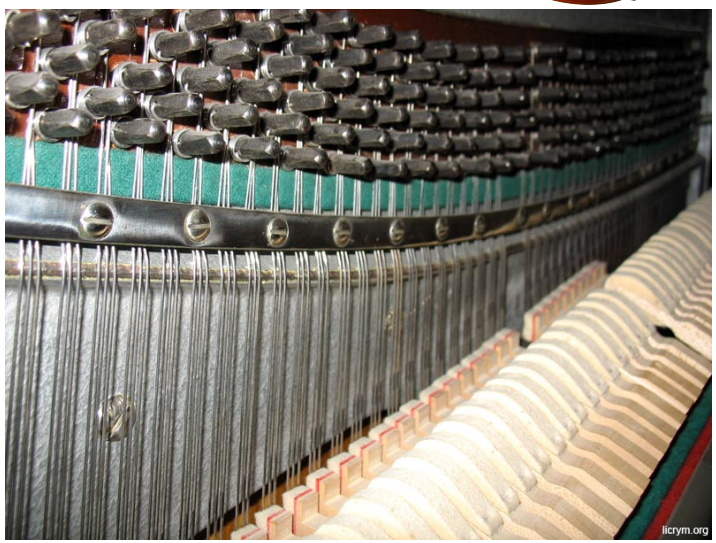
29 сентября 2014 г.

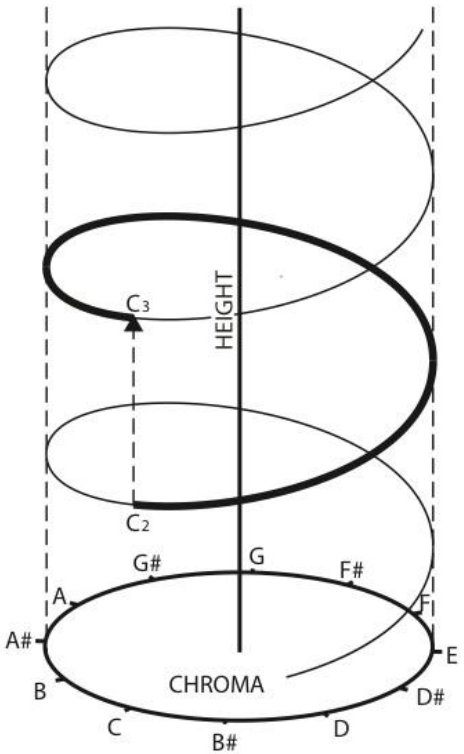
1. Анатомия музыки
   - Происхождение звуков
   - Разделение звуков
   - Эталонная нота
   - Ноты в музыкальных инструментах
   - Итоговое звучание
2. Цифровая звукозапись
   - Дискретизация и квантование
   - Цифровые аудиоформаты
   - Общая схема
3. Музыкальная транскрипция
   - Постановка задачи и мотивация
   - Преобразования Фурье
   - В чем же подвох?
4. Обзор существующих подходов
   - Предобработка сигнала
   - Подходы к распознаванию
   - Сравнение результатов

Звук — распространение в виде упругих волн механических колебаний в твёрдой, жидкой или газообразной среде. Обычный человек способен слышать звуковые колебания в диапазоне частот от 16 Гц до 20 кГц.



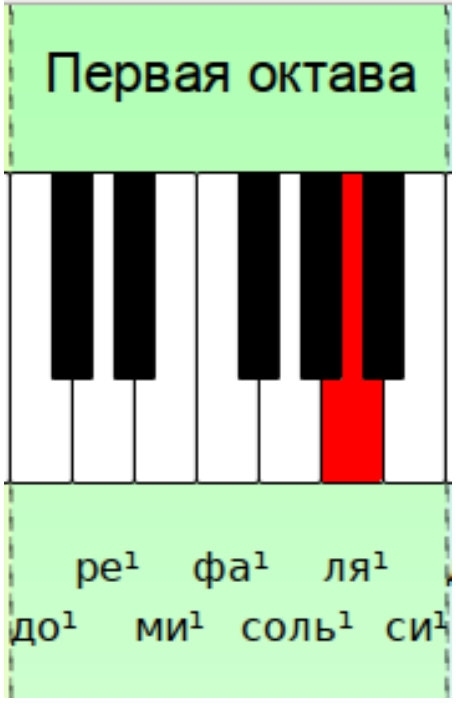
В музыке звук обычно получается колебаниями частей инструмента, исключение — духовые инструменты.





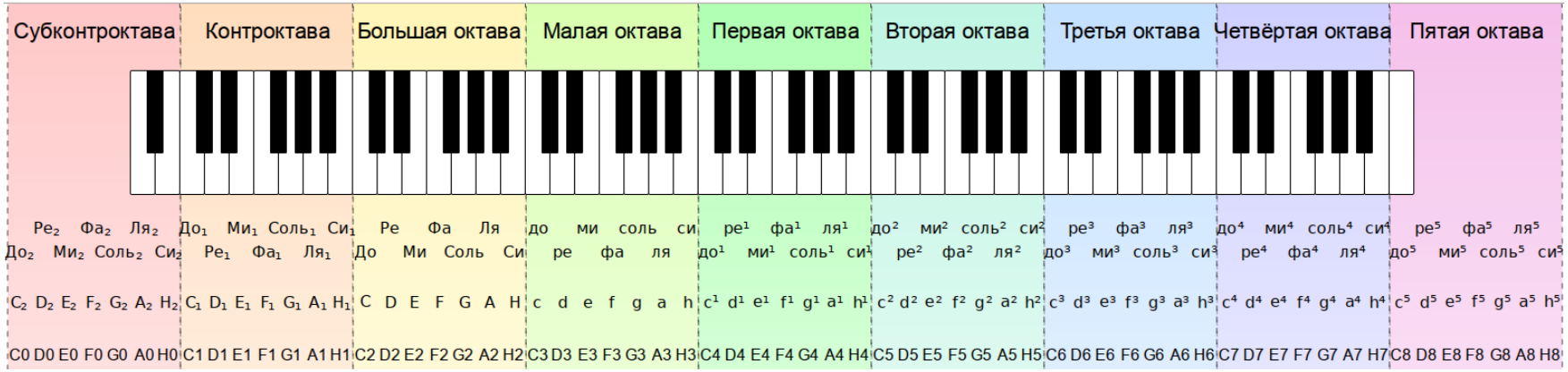
Музыкальные звуки, частота которых отличается в два раза, воспринимаются на слух как очень похожие, как повторение одного звука на разной высоте. Это явление называется октавным сходством звуков.
На основе этого весь диапазон частот используемых в музыке звуков делится на участки, называемые октавами, при этом частота звуков в каждой последующей октаве будет в два раза выше чем в предыдущей, а схожие звуки получают одинаковые названия ступеней.

За эталон частоты ноты берётся нота, частота которой должна быть равной 440 Гц. Это нота ля первой октавы.



Анатомия музыки

Ноты в музыкальных инструментах



Клавиатура фортепиано или рояля состоит из 88 клавиш — им соответствуют 7 полных октав и еще 4 ноты. Для многих инструментов такого соответствия не предусмотрено — при игре на скрипке можно надеяться только на свой слух.

Каждая октава делится на математически равные интервалы, в наиболее типичном случае — на двенадцать полутонов (каждый из которых равен $1 : \sqrt[12]{2}$).
Можно математически вычислить частоты для всего звукоряда, пользуясь формулой:

$$\mathrm{f}(\mathrm{i}) = \mathrm{f}_0 \cdot 2^{\mathrm{i}/12}$$

где f_0 — частота камертона (Ля, 440 Hz), а i — количество полутонов в интервале от искомого звука к эталону f_0.
К примеру, можно вычислить частоту звука на тон (2 полутона) ниже от камертона Ля — ноты соль:

$$\mathrm{f}(-2) = 440\,\mathrm{Hz} \cdot 2^{-2/12} \approx 391,995\,\mathrm{Hz}$$

Звукам, извлекаемым из фортепиано, в итоге соответствуют частоты от приблизительно 16 Гц до 15800 Гц.

Таблица соответствия нот частотам [править | править вики-текст]

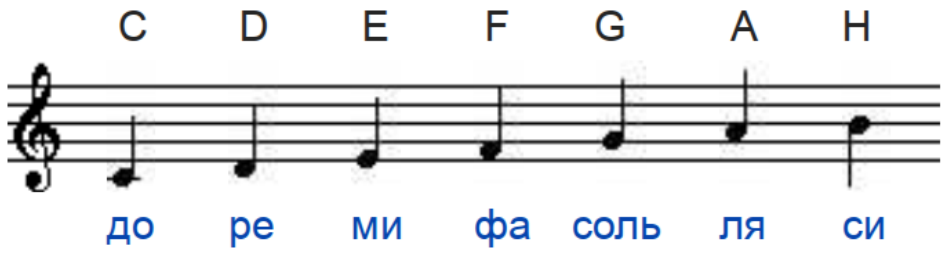
Частоты в герцах (интервал от средней До в полутонах)										
Октава → Нота ↓	Суб-контр	Контр	Большая	Малая	1	2	3	4	5	6
C	16,352 (−48)	32,703 (−36)	65,406 (−24)	130,81 (−12)	261,63 (0)	523,25 (+12)	1046,5 (+24)	2093,0 (+36)	4186,0 (+48)	8372,0 (+60)
C# / D♭	17,324 (−47)	34,648 (−35)	69,296 (−23)	138,59 (−11)	277,18 (+1)	554,37 (+13)	1108,7 (+25)	2217,5 (+37)	4434,9 (+49)	8869,8 (+61)
D	18,354 (−46)	36,708 (−34)	73,416 (−22)	146,83 (−10)	293,66 (+2)	587,33 (+14)	1174,7 (+26)	2349,3 (+38)	4698,6 (+50)	9397,3 (+62)
D# / E♭	19,445 (−45)	38,891 (−33)	77,782 (−21)	155,56 (−9)	311,13 (+3)	622,25 (+15)	1244,5 (+27)	2489,0 (+39)	4978,0 (+51)	9956,1 (+63)
E	20,602 (−44)	41,203 (−32)	82,407 (−20)	164,81 (−8)	329,63 (+4)	659,26 (+16)	1318,5 (+28)	2637,0 (+40)	5274,0 (+52)	10548 (+64)
F	21,827 (−43)	43,654 (−31)	87,307 (−19)	174,61 (−7)	349,23 (+5)	698,46 (+17)	1396,9 (+29)	2793,8 (+41)	5587,7 (+53)	11175 (+65)
F# / G♭	23,125 (−42)	46,249 (−30)	92,499 (−18)	185,00 (−6)	369,99 (+6)	739,99 (+18)	1480,0 (+30)	2960,0 (+42)	5919,9 (+54)	11840 (+66)
G	24,500 (−41)	48,999 (−29)	97,999 (−17)	196,00 (−5)	392,00 (+7)	783,99 (+19)	1568,0 (+31)	3136,0 (+43)	6271,9 (+55)	12544 (+67)
G# / A♭	25,957 (−40)	51,913 (−28)	103,83 (−16)	207,65 (−4)	415,30 (+8)	830,61 (+20)	1661,2 (+32)	3322,4 (+44)	6644,9 (+56)	13290 (+68)
A	27,500 (−39)	55,000 (−27)	110,00 (−15)	220,00 (−3)	440,00 (+9)	880,00 (+21)	1760,0 (+33)	3520,0 (+45)	7040,0 (+57)	14080 (+69)
A# / H♭	29,135 (−38)	58,270 (−26)	116,54 (−14)	233,08 (−2)	466,16 (+10)	932,33 (+22)	1864,7 (+34)	3729,3 (+46)	7458,6 (+58)	14917 (+70)
H	30,868 (−37)	61,735 (−25)	123,47 (−13)	246,94 (−1)	493,88 (+11)	987,77 (+23)	1975,5 (+35)	3951,1 (+47)	7902,1 (+59)	15804 (+71)

Примечание: Иногда нота Си♭ обозначается как «B» вместо «H♭».

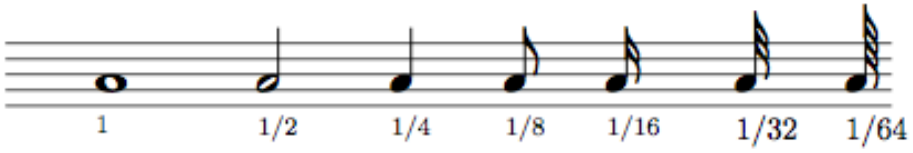
Анатомия музыки

Ноты в музыкальных инструментах

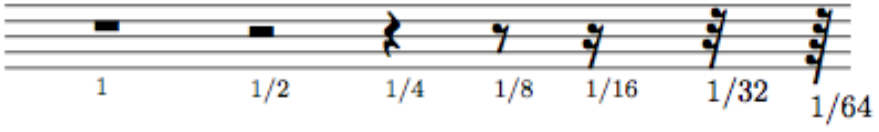
Для записи используются нотные знаки на нотном стане:



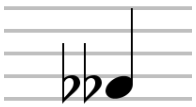
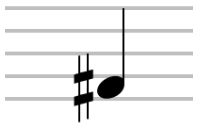
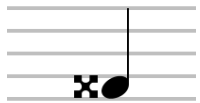
Разные длительности нот:



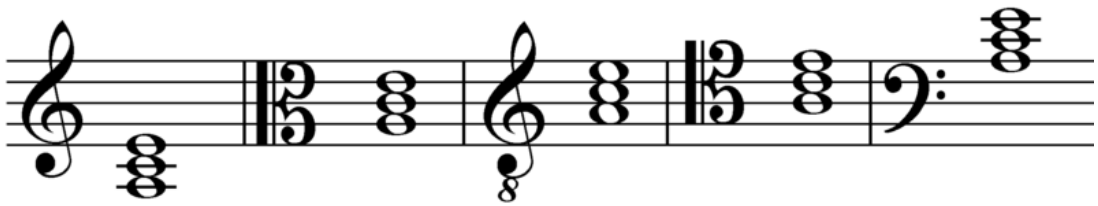
Изображения пауз:



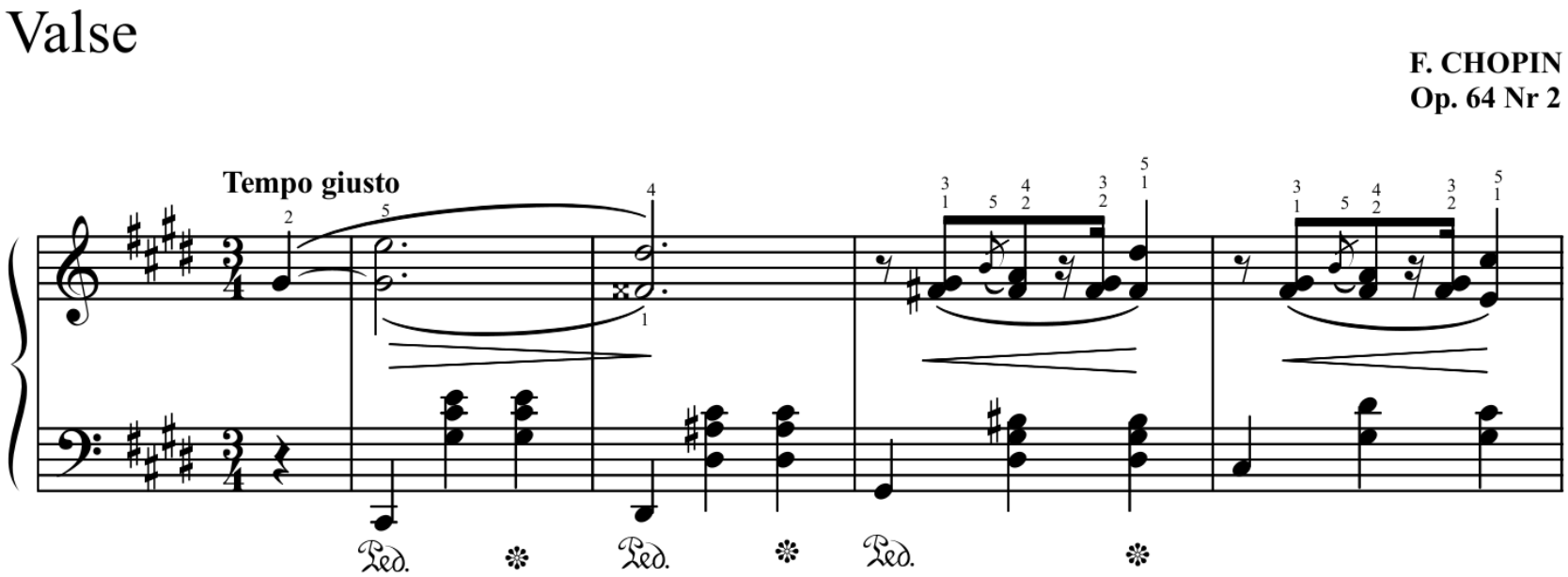
Знаки альтерации:

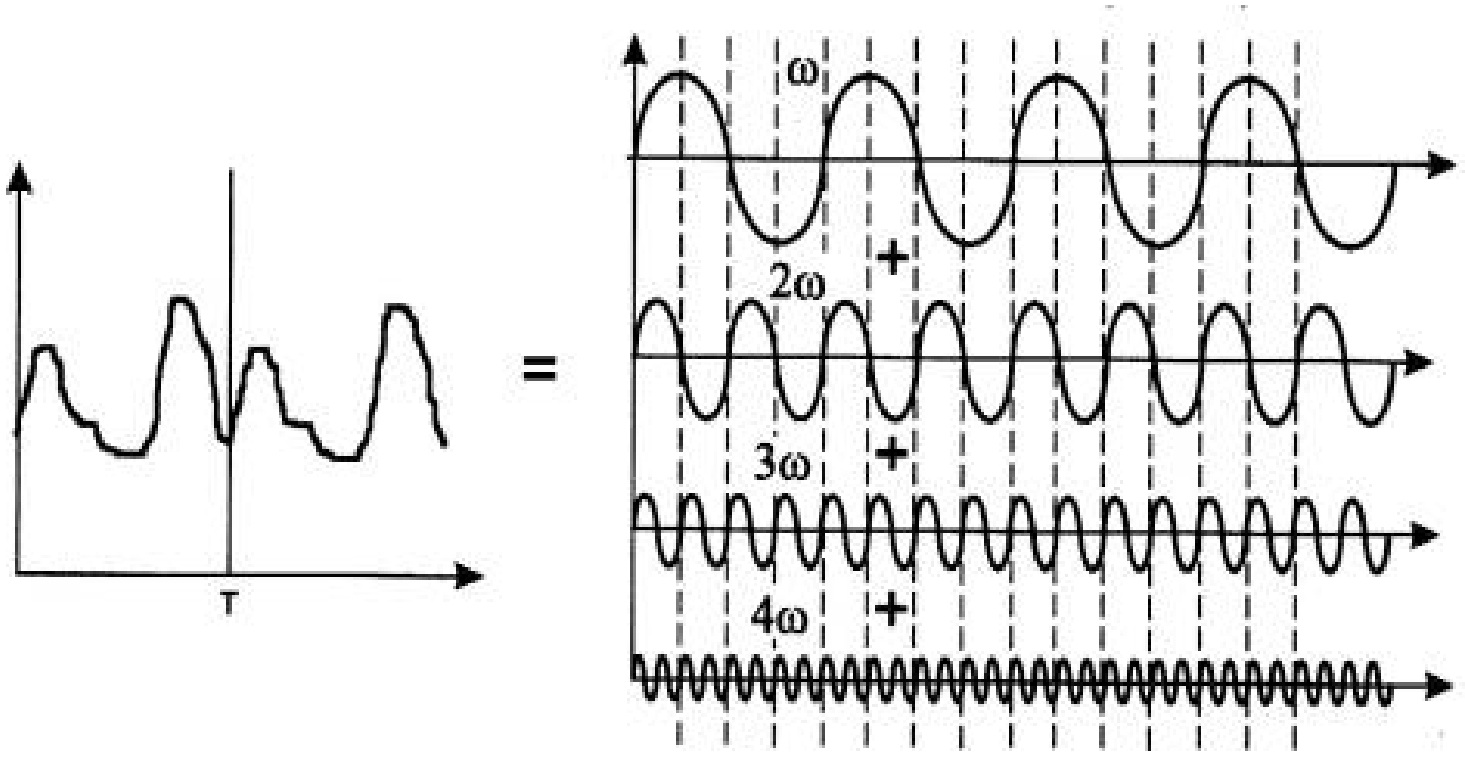
Пример *трезвучия* в разных записях (ключах):



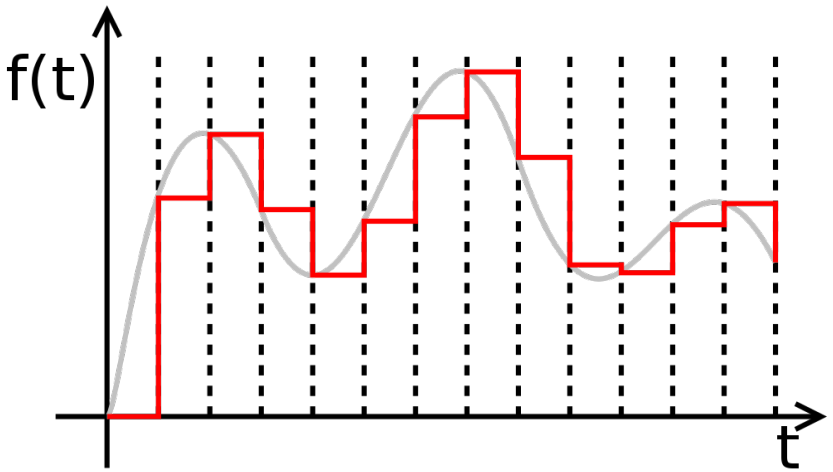
Пример нотной записи:



Итоговое музыкальное звучание получается наложением всех извлекаемых из инструмента звуков.



Для сохранения звука в компьютерах используется т.н. дискретизация — мы сохраняем значение колебания звука через равные промежутки времени. Количество таких измерений в секунду — частота дискретизации. Количество возможных измеренных значений — разрядность квантования.



Характеристика для Audio CD — 44100 Гц / 16 бит.

Формат файла определяет структуру и особенности представления звуковых данных при хранении на компьютере. Для устранения избыточности данных используются аудиокодеки, при помощи которых производится сжатие аудиоданных. Выделяют три группы звуковых форматов аудиофайлов:

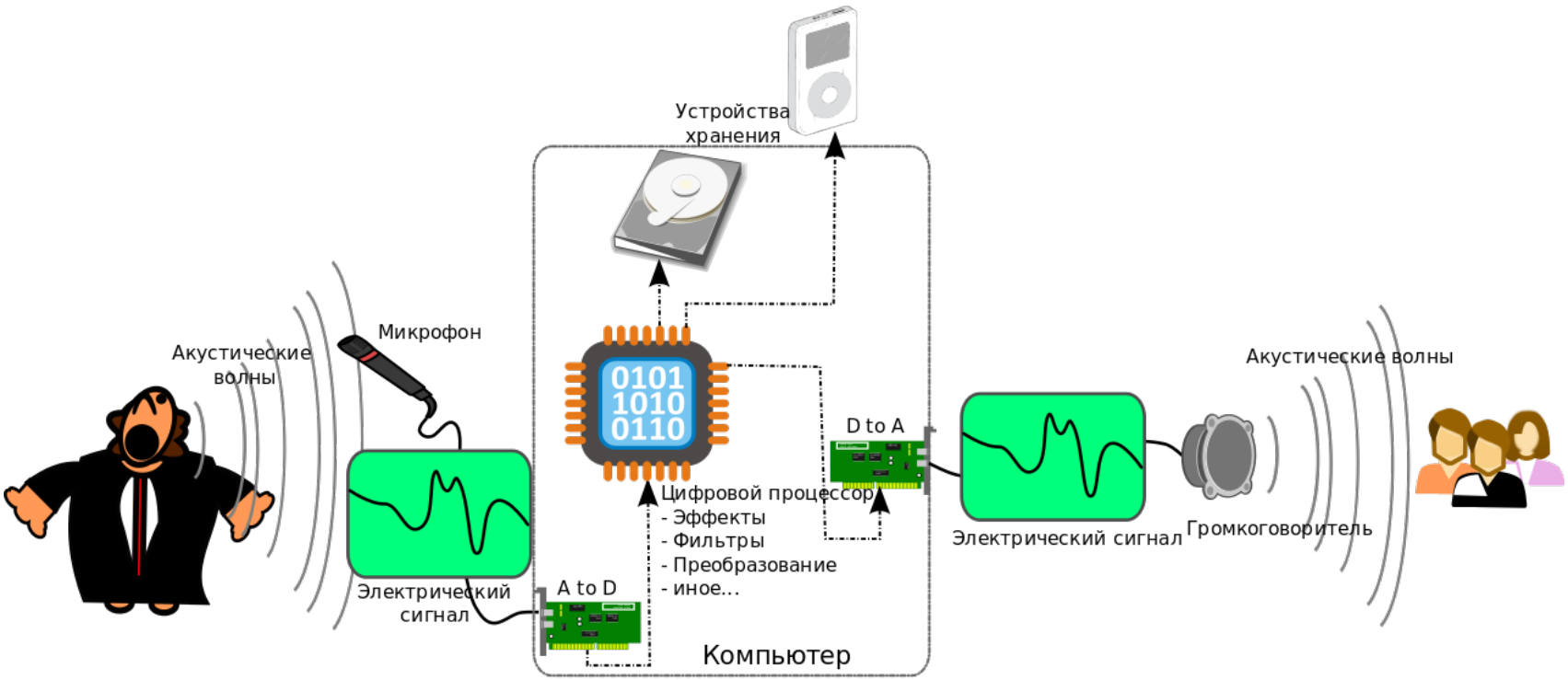
- форматы без сжатия (WAV, AIFF)
- форматы со сжатием без потерь (APE, FLAC)
- форматы с применением сжатия с потерями (mp3, ogg)

Некоторые виды цифровых аудиформатов в сравнении:

Название формата	Квантование, бит	Частота дискретизации, кГц	Число каналов	Величина потока данных с диска, кбит/с	Степень сжатия/упаковки
CD	16	44,1	2	1411,2	1:1 без потерь
Dolby Digital (AC3)	16-24	48	6	до 640	~12:1 с потерями
DTS	20-24	48; 96	до 8	до 1536	3:1 с потерями
DVD-Audio	16; 20; 24	44,1; 48; 88,2; 96	6	6912	1:1 без потерь
DVD-Audio	16; 20; 24	176,4; 192	2	4608	1:1 без потерь
MP3	16-24	до 48	2	до 320	~11:1 с потерями
AAC	16-24	до 96	до 48	до 512	с потерями
AAC+ (SBR)	16-24	до 48	2	до 320	с потерями
Ogg Vorbis	до 32	до 192	до 255	до 500	с потерями
WMA	до 24	до 96	до 8	до 768	2:1, есть версия без потерь

Цифровая звукозапись

Общая схема



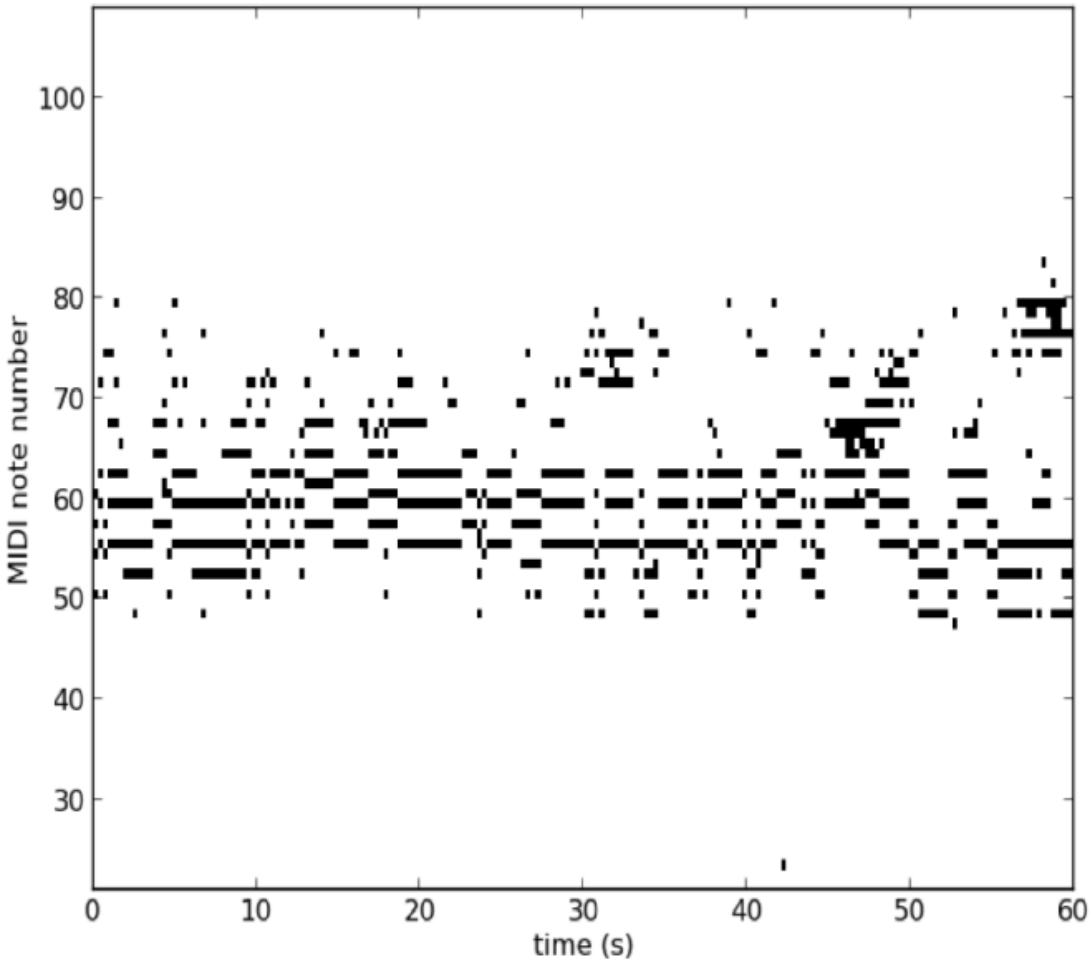
Исследуются полифонические записи - т.е. записи, на которых воспроизводятся множество нот инстурмента одновременно.
При этом решаются следующие задачи:

- pitch detection — определяем, когда какие ноты воспроизводились исполнителем
- beat detection — какой ритм у музыки (как часто надо стучать ногой «в такт» музыке :)

Для чего?

- помощь при обучении, *составлении нот*
- определение жанра
- определение автора/композитора
- поиск плагиата в музыке

Хотелось бы получить что-то вроде этого:



Если звук в своем самом простом виде — синусоида, и музыка состоит из множества таких звуков, то сигнал можно разложить в сумму таких синусоид? Тригонометрический ряд Фурье — представление функции f с периодом τ в виде ряда

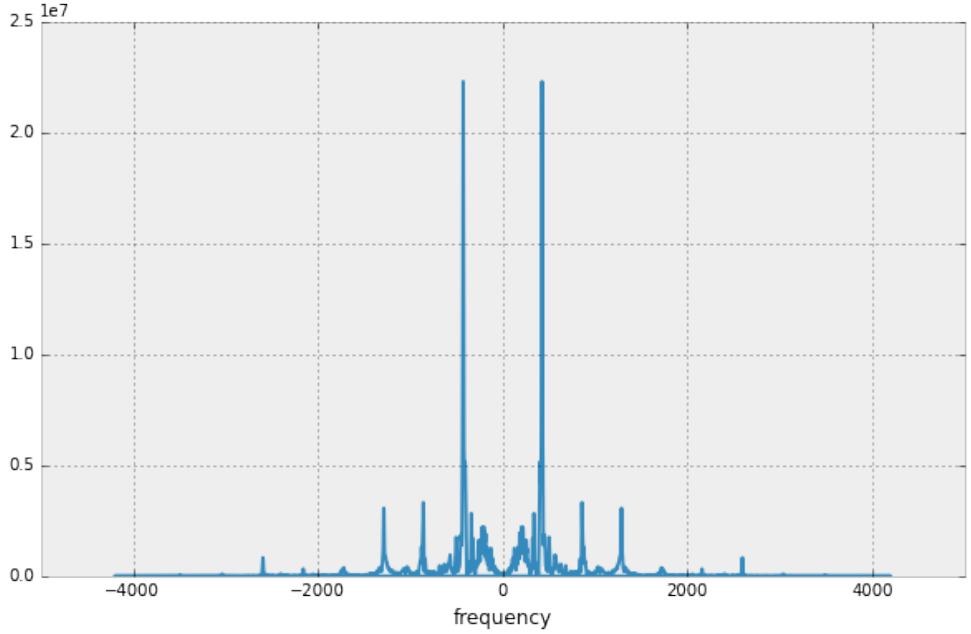
$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx).$$

Для дискретизованного сигнала будем использовать дискретное преобразование Фурье:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1.$$

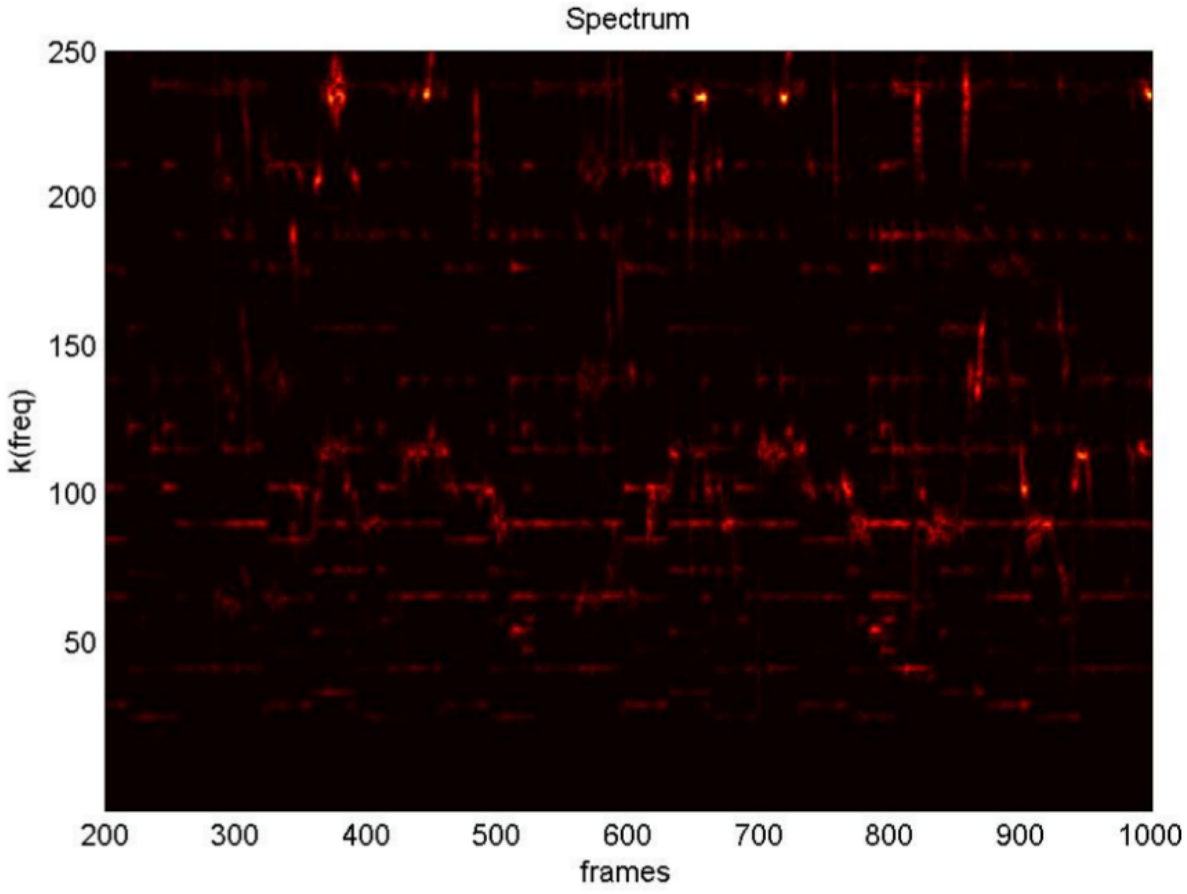
Быстрое преобразование фурье (FFT) — алгоритм для его быстрого подсчета.

На основе преобразования Фурье для малого промежутка времени можно также рассмотреть спектр частот, чтобы понять, какие частоты преобладают для этого промежутка.



(на рисунке спектр участка с звучащей с ля-бемоль 1-й октавы)

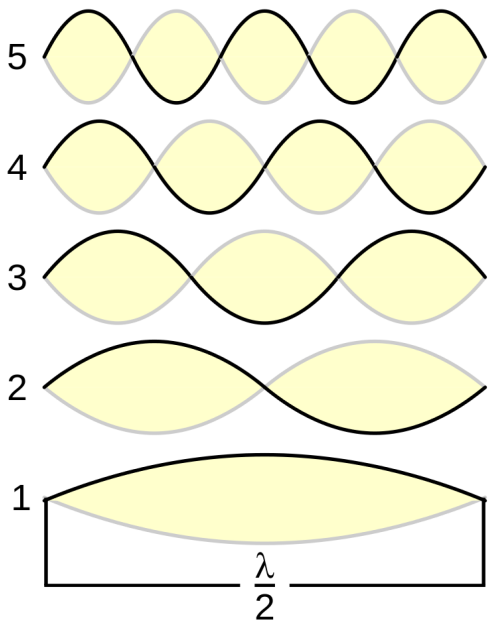
Все произведение можно представить в виде спектрограммы.



Музыкальная транскрипция

В чем же подвох?

Казалось бы - все просто, по спектрограмме ищем внезапно возросшие частоты наших синусоид и определяем, какие ноты для этого нажимались (потом объединяем в аккорды).
Но существует проблема:

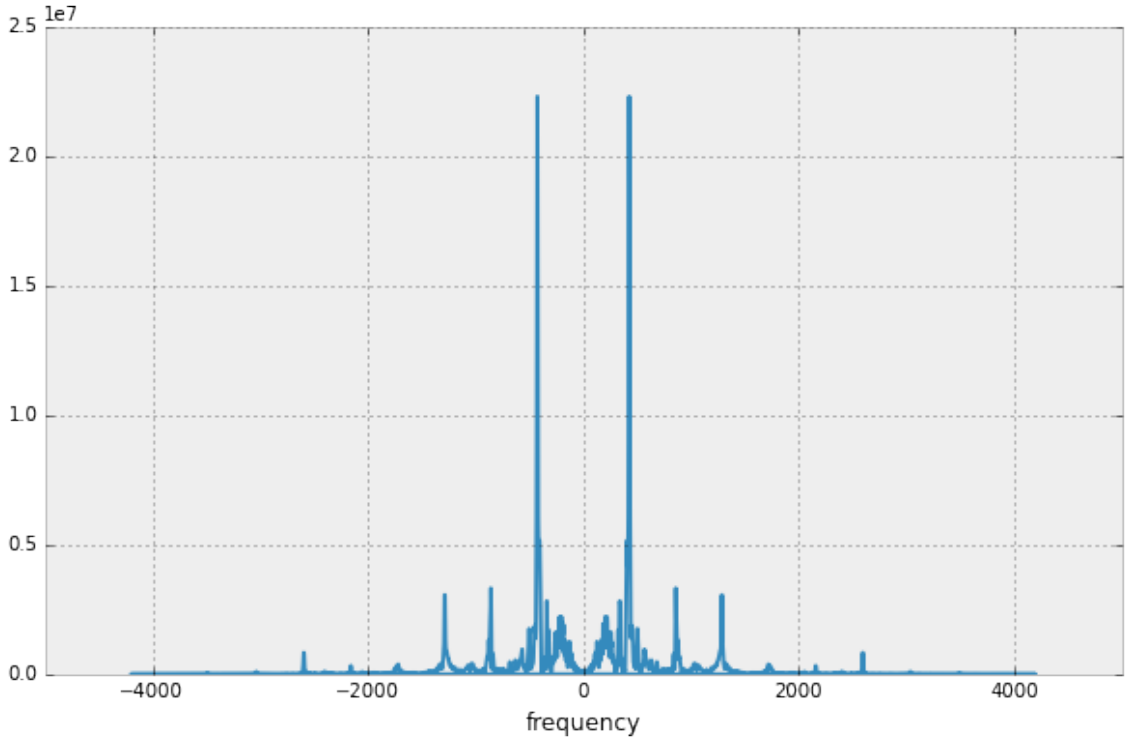


Обертон - колебания частей звучащего тела, которые создают звуки повышенной частоты в простом случае кратной начальной (гармонический обертон).
Начальные 10 обертонов прослушиваются по высоте и сливаются друг с другом в аккорды. Уже первый обертон доставит нам не мало хлопот — как отличить при игре октаву (интервал) от обертонов?
Помимо этого есть еще просто шум.

Музыкальная транскрипция

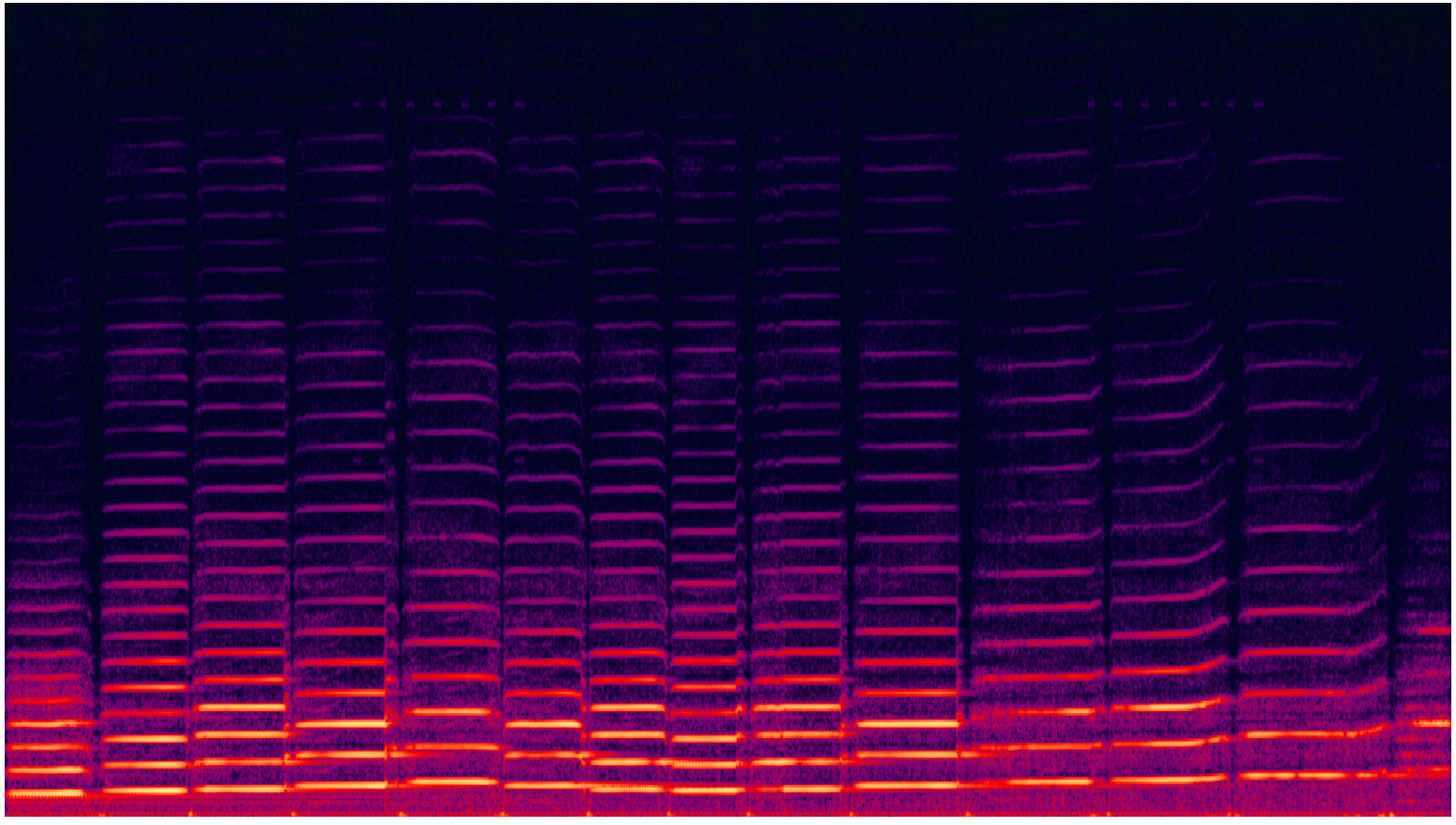
В чем же подвох?

Пример, рассмотренный ранее - звучит ля бемоль 1-й октавы.



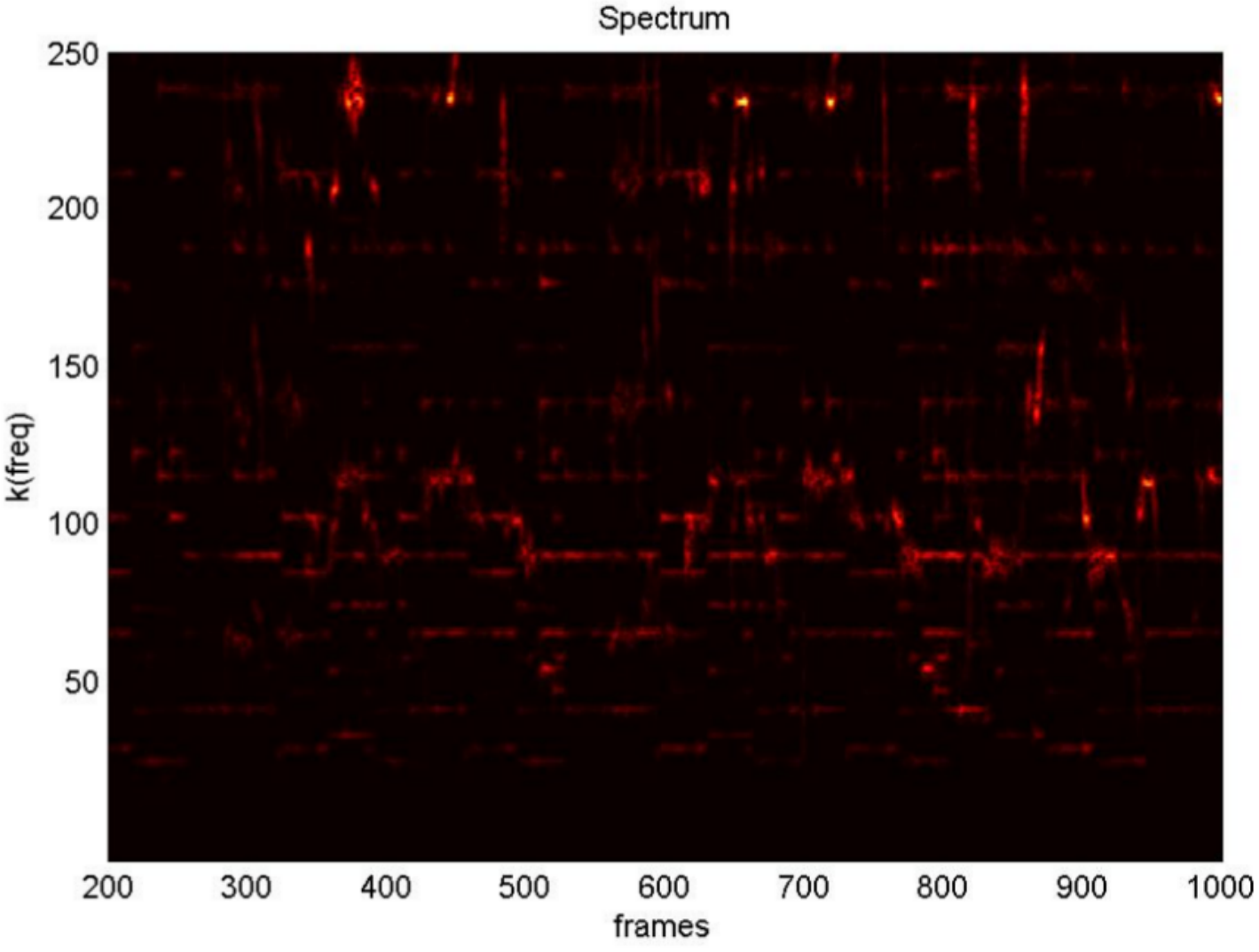
На графике заметны пики мощности, сотоветствующие первым двум обертонам нашего звука.

Пример — спектрограмма игры на скрипке (всего 4 струны).



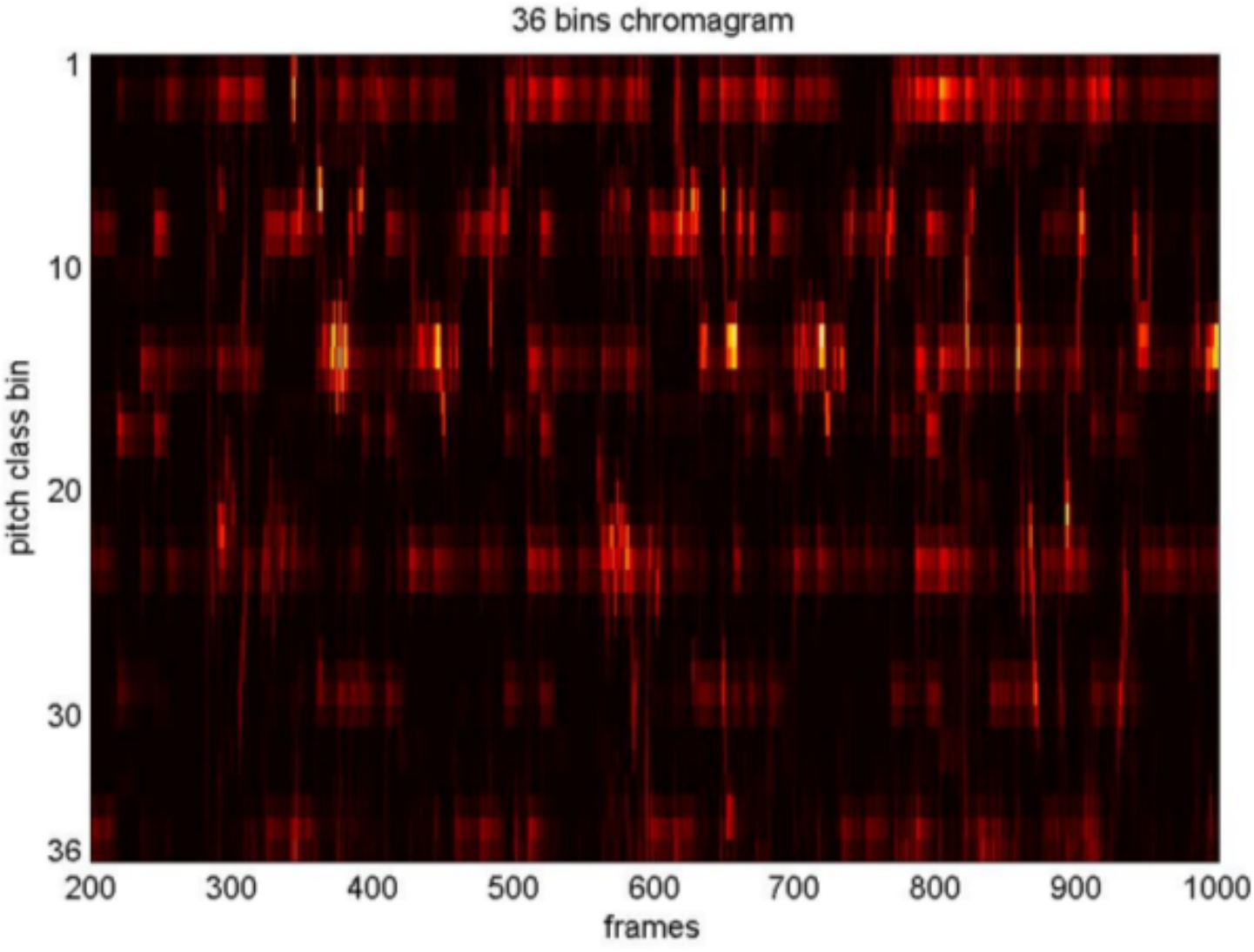
Обзор существующих подходов

Предобработка сигнала



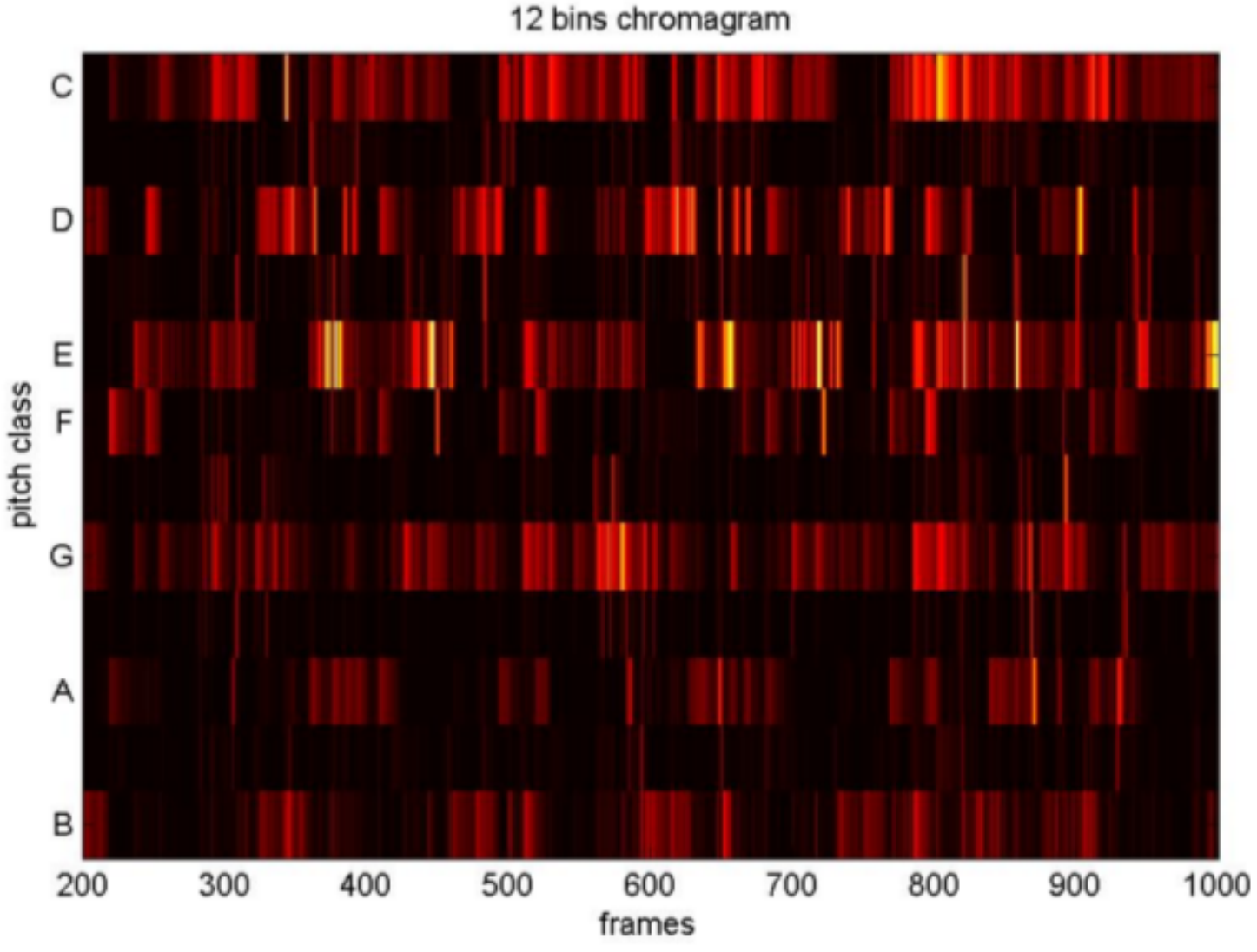
Обзор существующих подходов

Предобработка сигнала



Обзор существующих подходов

Предобработка сигнала



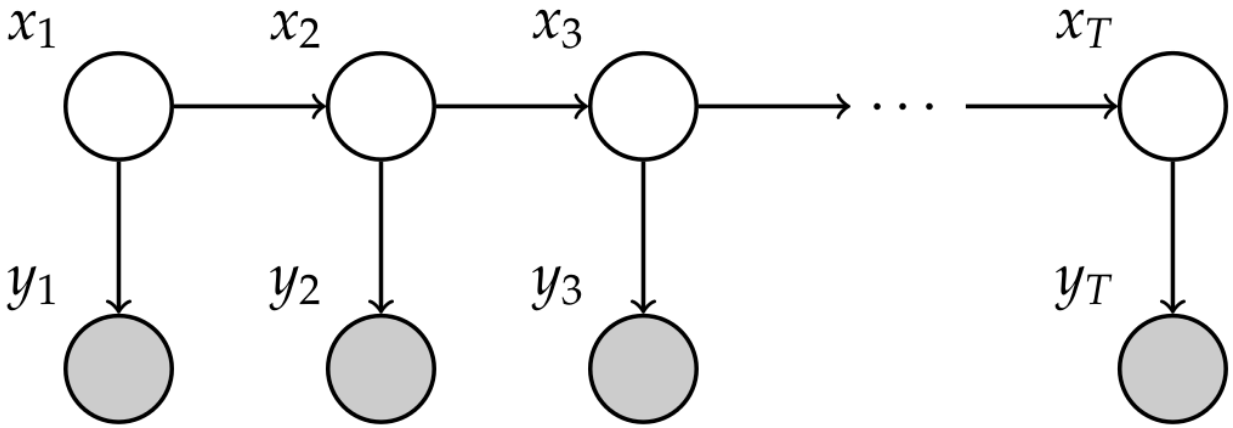
- E. Poliner, P. W. Ellis — A Discriminative Model for Polyphonic Piano Transcription (2006)
- M. Marolt — A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music (2004)
- M. Ryynänen, A. Klapuri — Polyphonic Music Transcription Using Note Event Modeling for MIREX 2008
- A. Zalani, A. Mittal — Polyphonic Music Transcription: A Deep Learning Approach (2014)

Poliner, Ellis (2006)

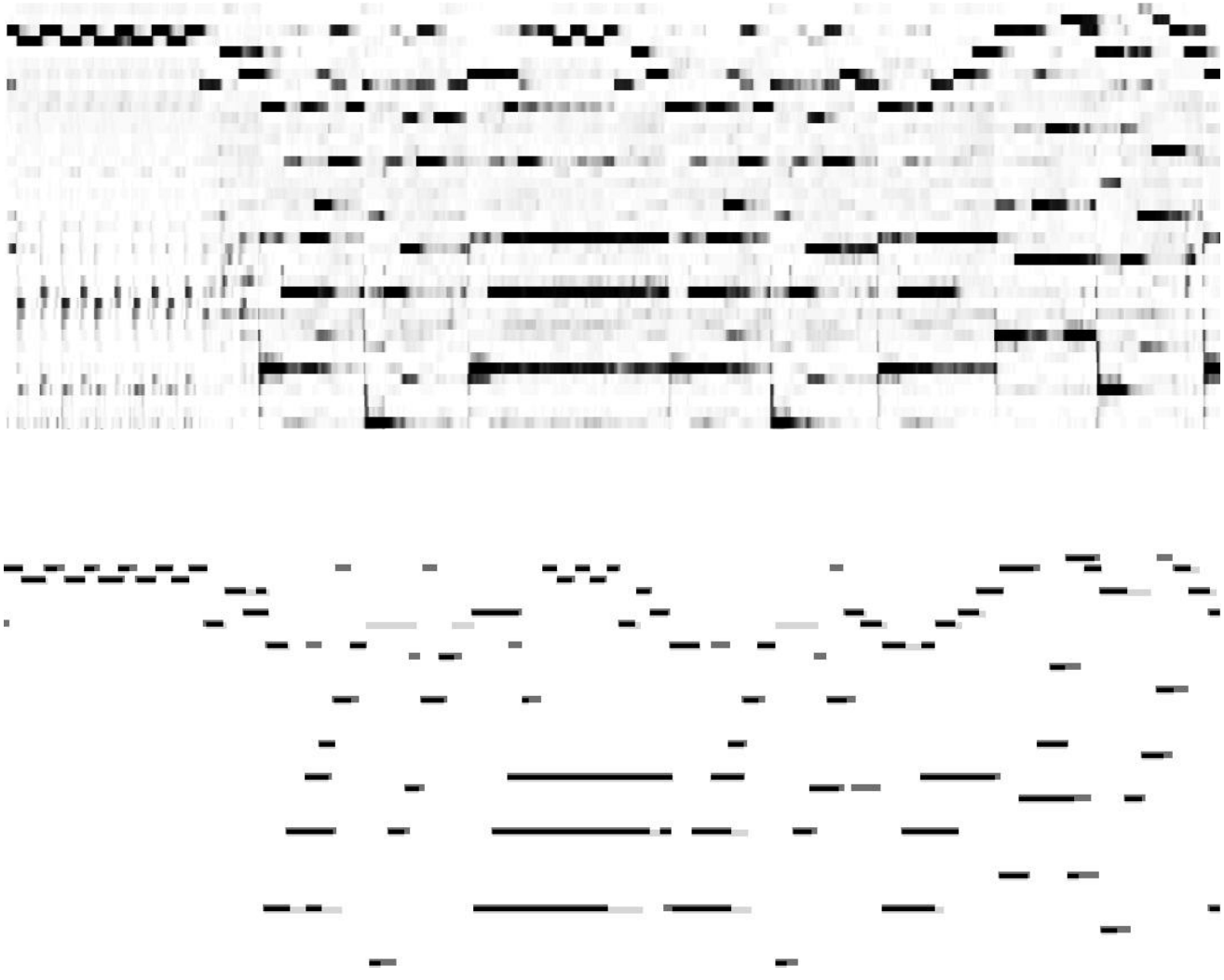
- стандартный спектральный анализ
- 87 SVM-RBF-классификаторов One-vs-All
- пост-процессинг скрытыми марковскими моделями:

$$\prod_t p(c_t|q_t)p(q_t|q_{t-1}) \to \max$$

$$p(q_t|x_t) \propto p(x_t|q_t)p(q_t)$$



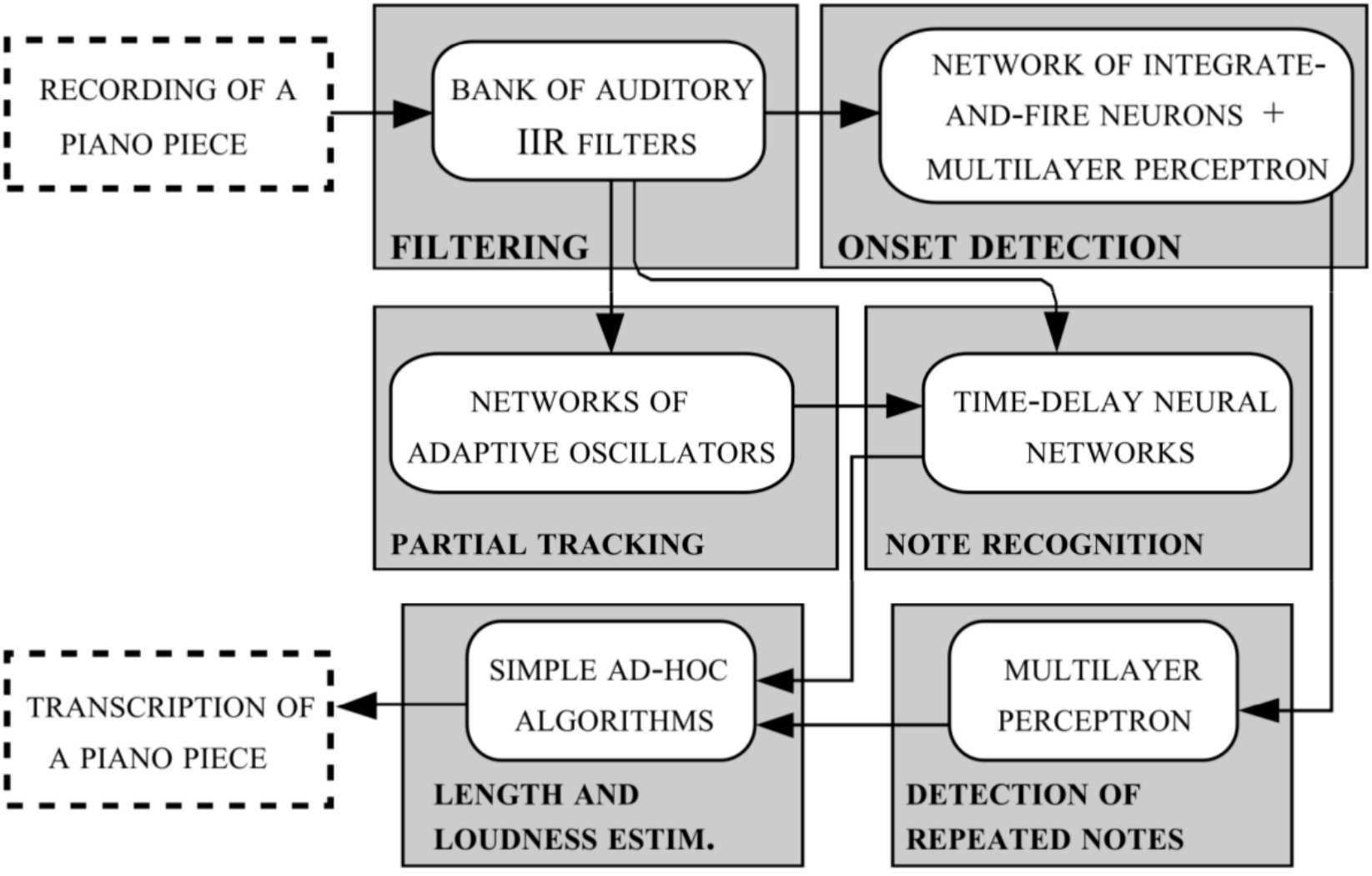
Poliner, Ellis (2006)



Marolt (2004)

- используются нейросети и их обобщения
- вводятся «адаптивные осцилляторы» для определения частот звуков
- с помощью особой синхронизации осцилляторов удается учитывать обертоны в музыке
- объединяя такие осцилляторы в сети, получем систему для отслеживания целых групп обертонов, т.е. фактически получаем необходимую транскрипцию

Marolt (2004)


RECORDING OF A PIANO PIECE
BANK OF AUDITORY IIR FILTERS
FILTERING
NETWORK OF INTEGRATE-AND-FIRE NEURONS + MULTILAYER PERCEPTRON
ONSET DETECTION
NETWORKS OF ADAPTIVE OSCILLATORS
PARTIAL TRACKING
TIME-DELAY NEURAL NETWORKS
NOTE RECOGNITION
TRANSCRIPTION OF A PIANO PIECE
SIMPLE AD-HOC ALGORITHMS
LENGTH AND LOUDNESS ESTIM.
MULTILAYER PERCEPTRON
DETECTION OF REPEATED NOTES

Marolt (2004)

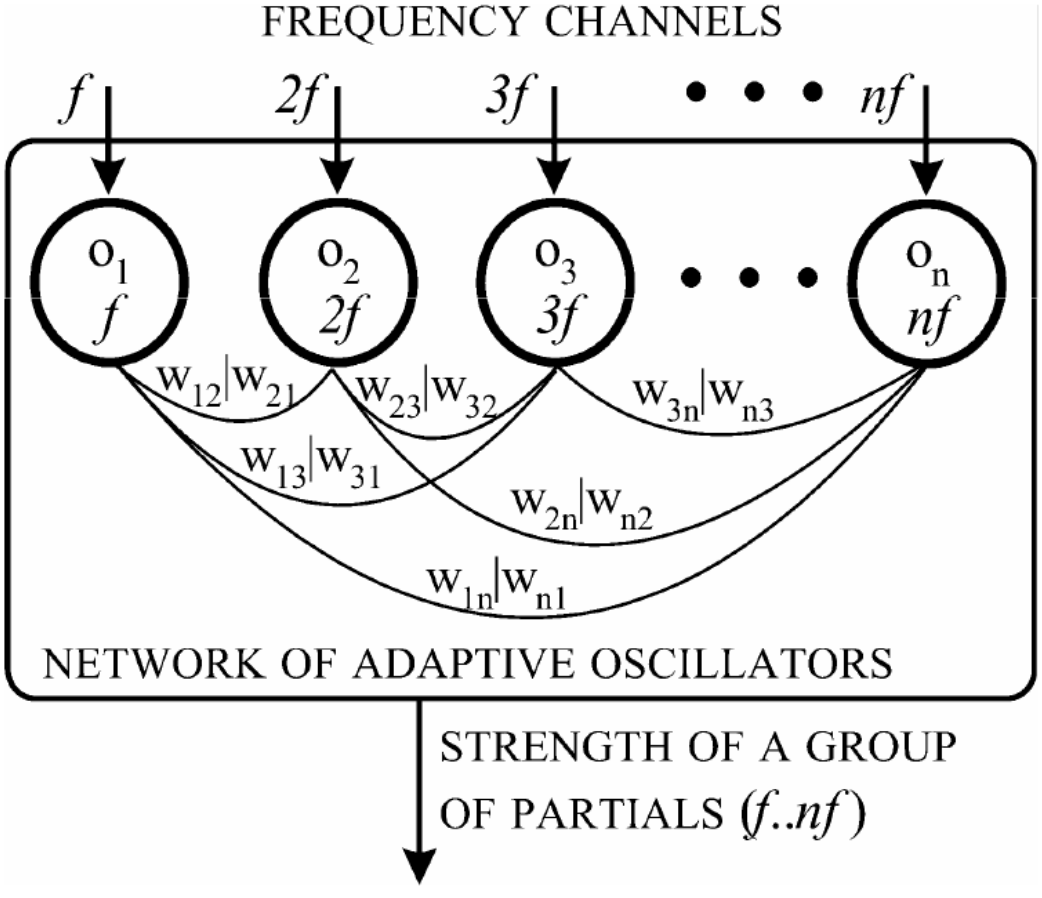


Fig. 3. A network of adaptive oscillators.

Marolt (2004)

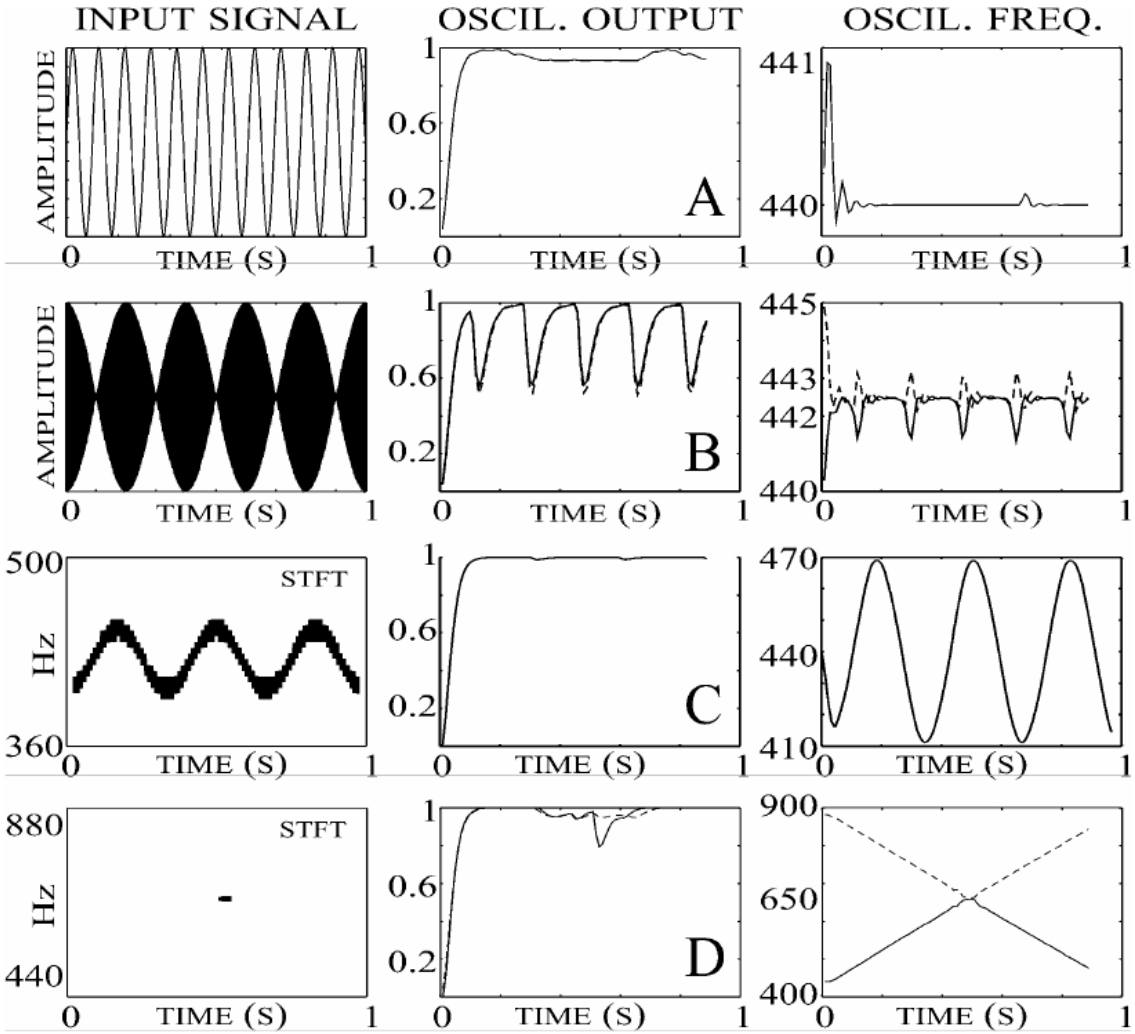


Fig. 2. Partial tracking with adaptive oscillators.

Ryynänen, Klapuri (2005)

- первичная обработка — multiple-F0 estimator
- HMM для нот, модель тишины и музыкологическая модель
- музыкологическая модель отвечает за поиск переходов между HMM и моделью тишины

Ryynänen, Klapuri (2005)

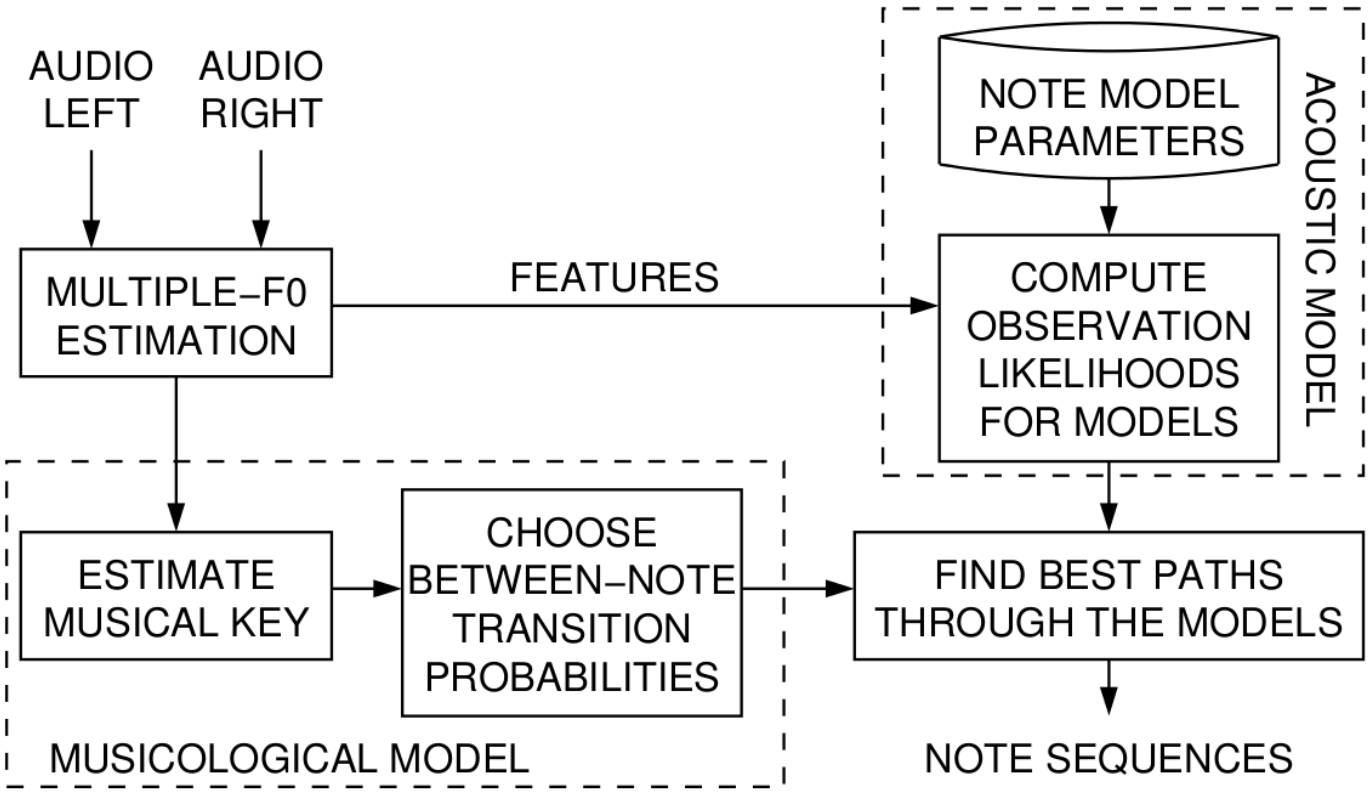


Figure 1. A block diagram of the transcription method.

Ryynänen, Klapuri (2005)

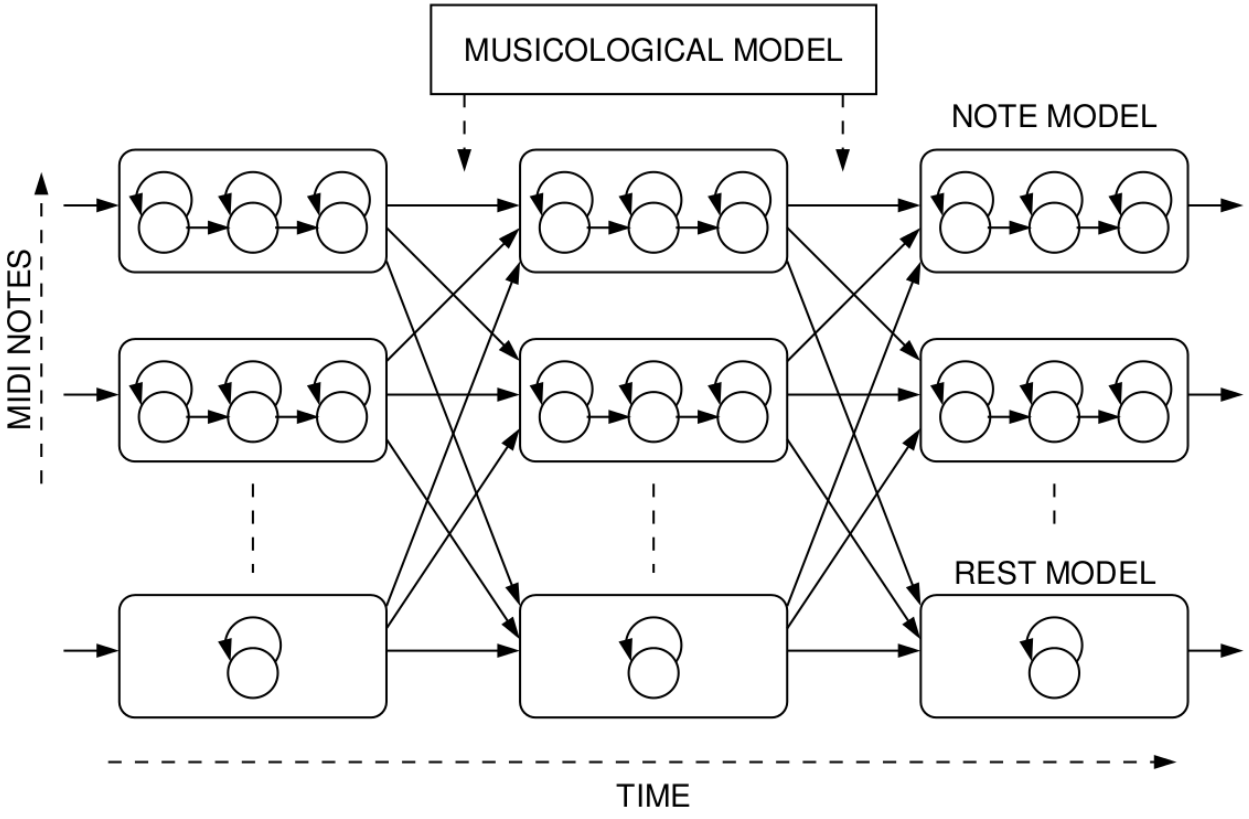
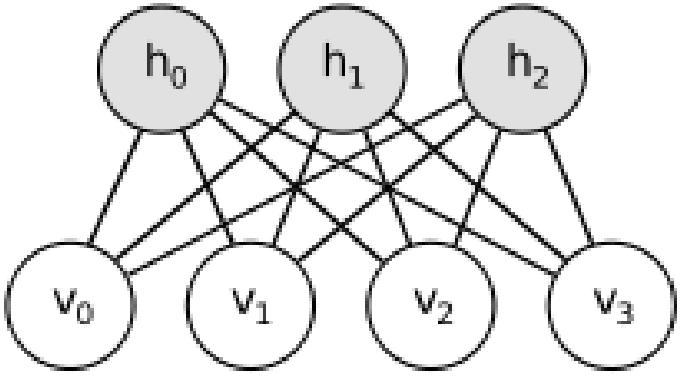
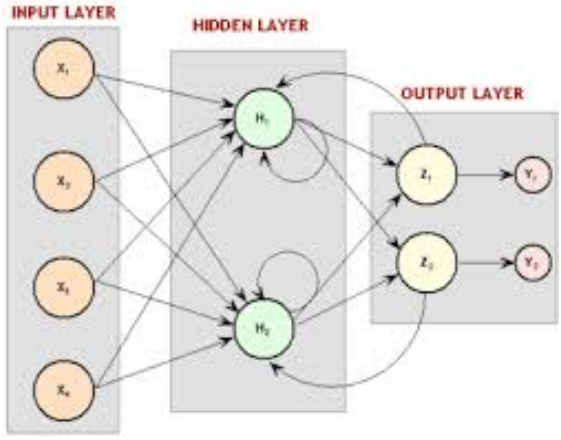


Figure 2. The network of note models and the rest model.

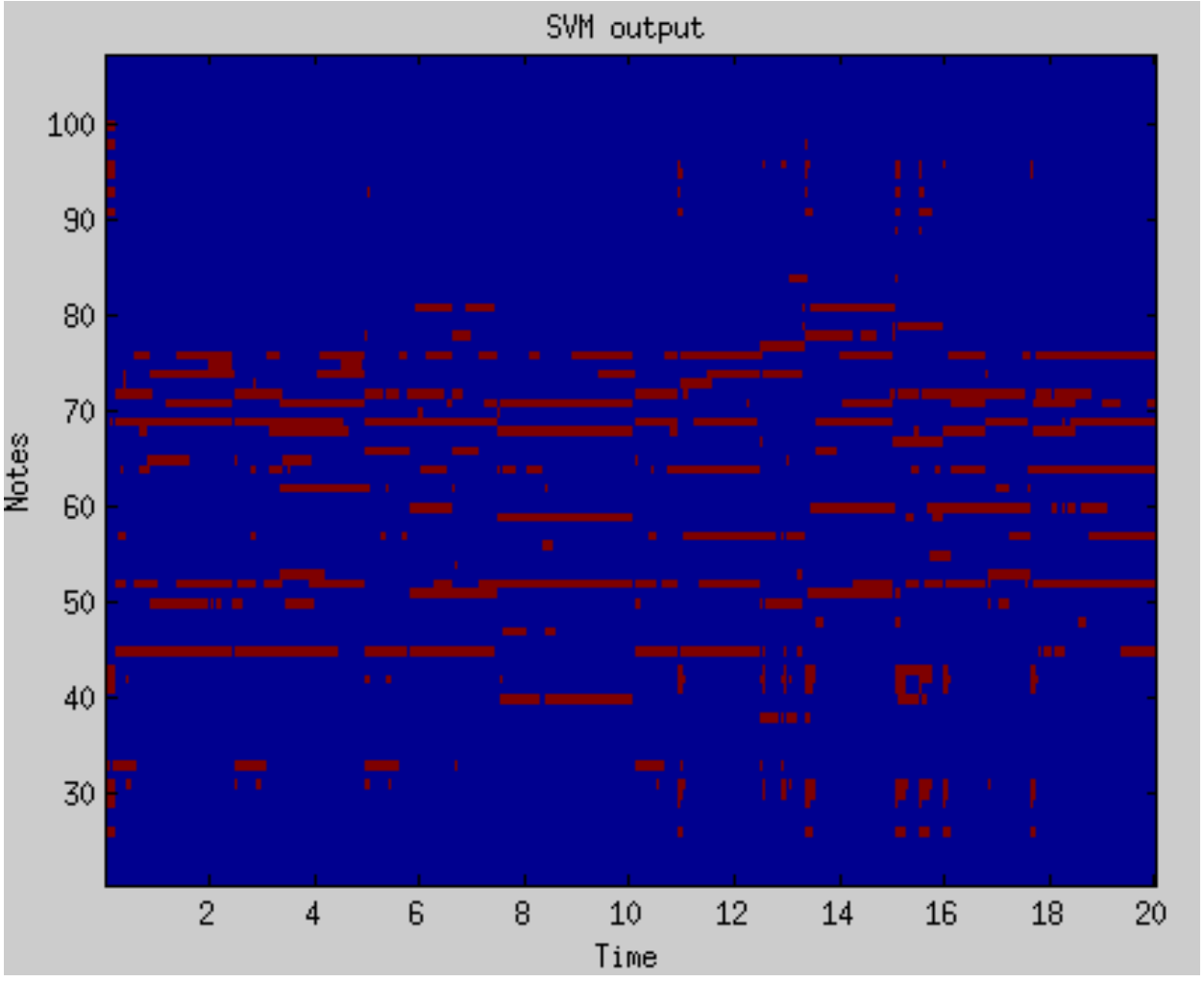
Zalani, Mittal (2014)

- RNN-RBM для обучения признаков
- STFT тоже учитывается как признаки
- 88 SVM-классификаторов One-vs-All
- пост-процессинг скрытыми марковскими моделями

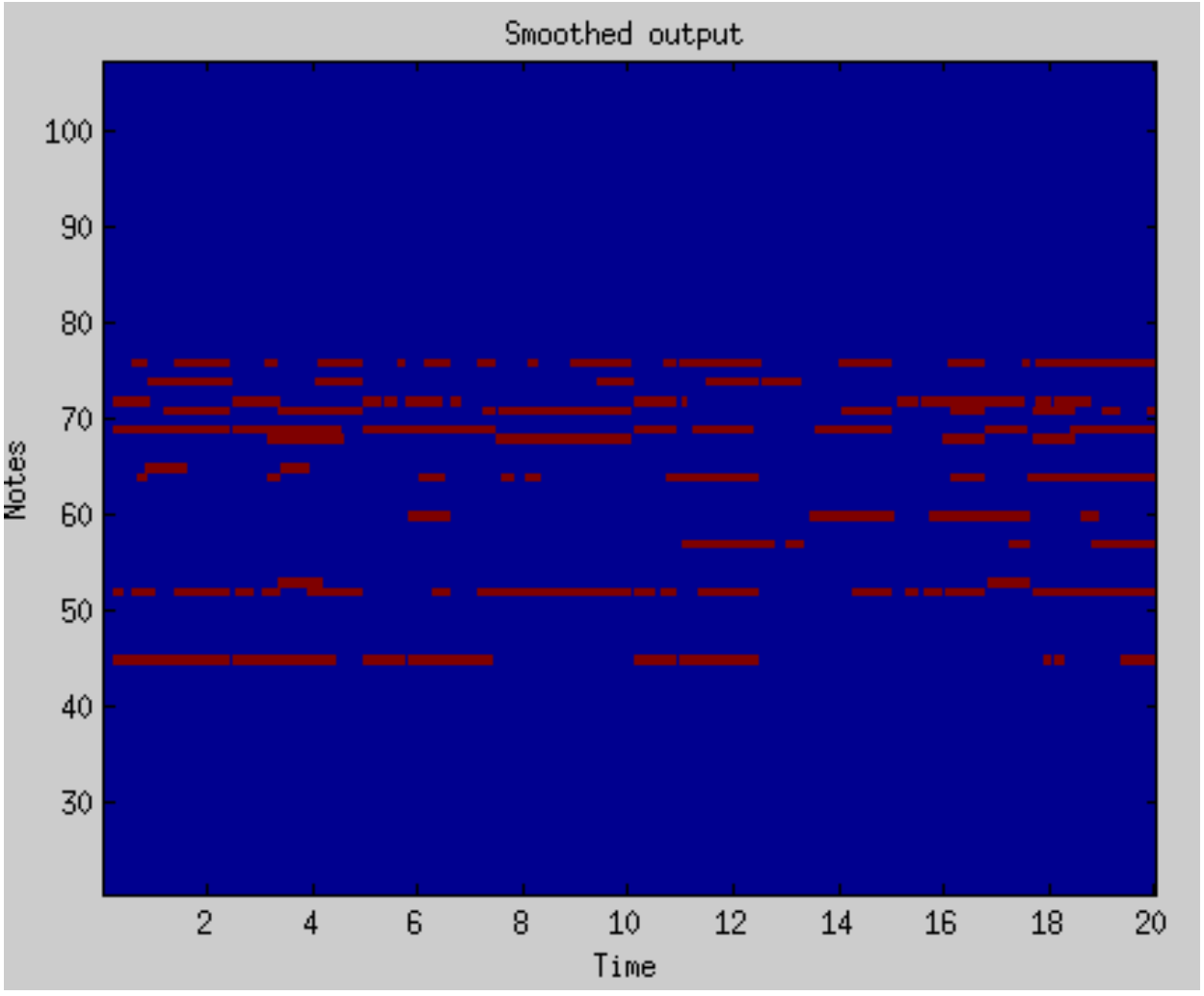




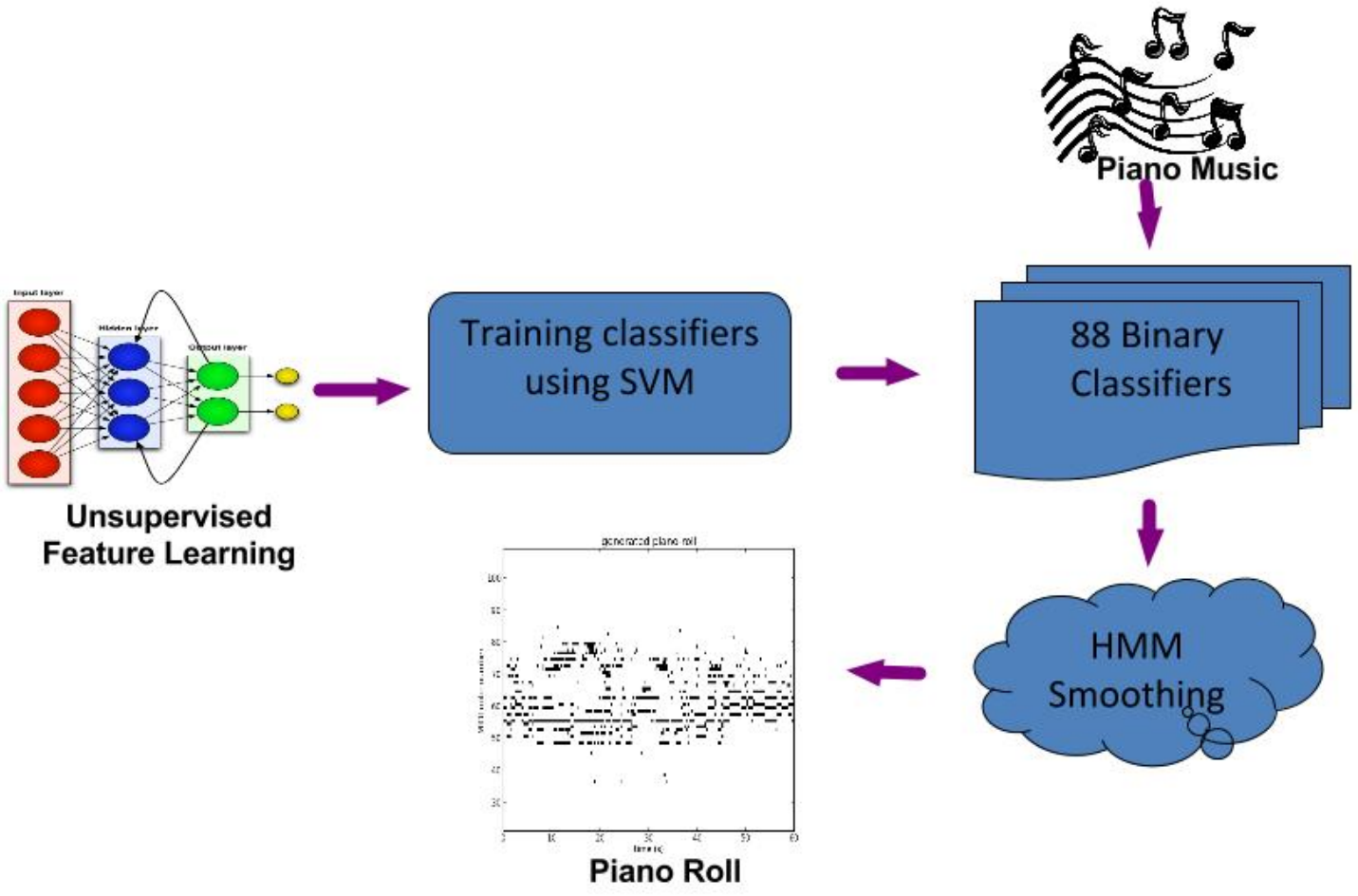
Zalani, Mittal (2014)



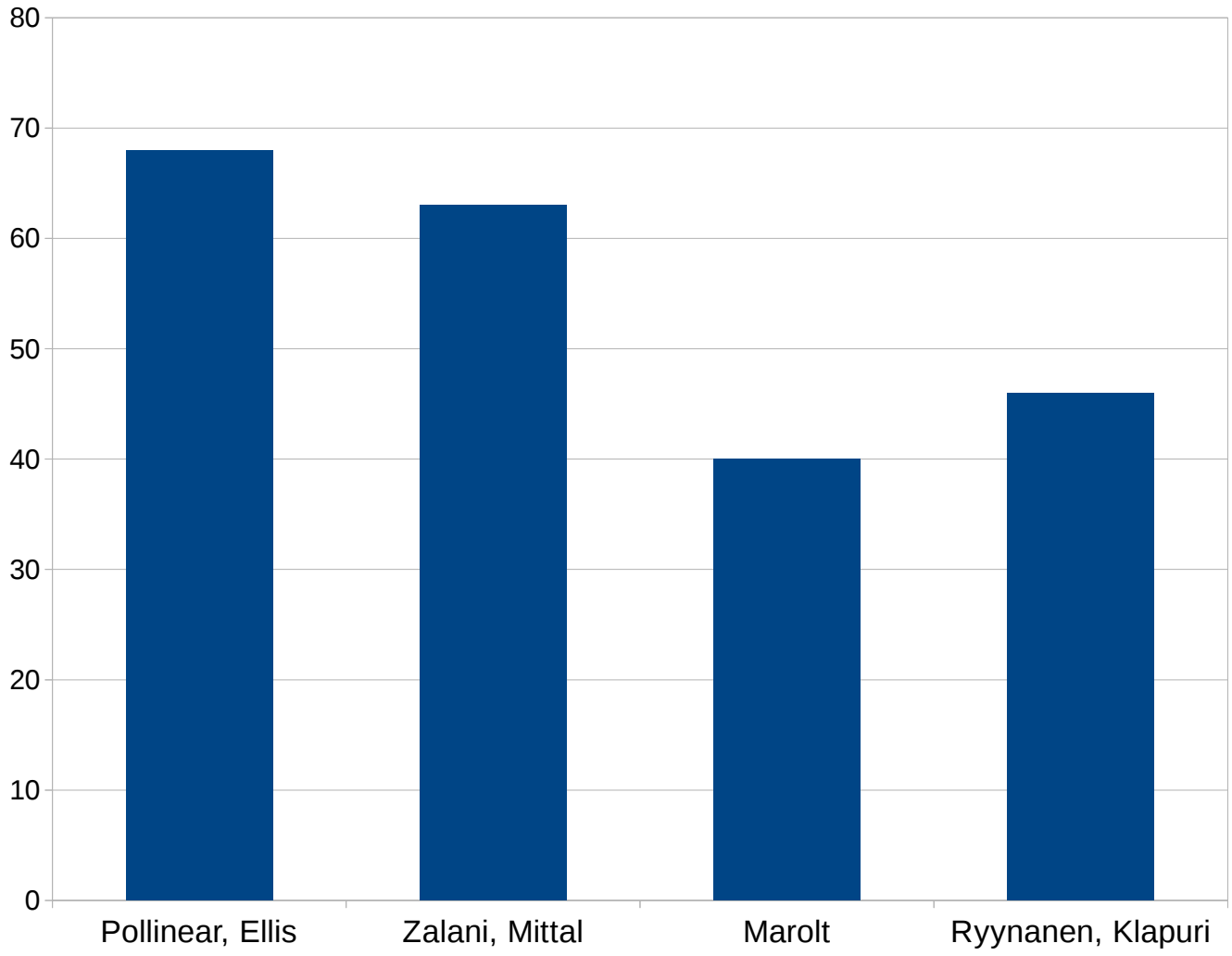
Zalani, Mittal (2014)



Zalani, Mittal (2014)



Точность распознавания (сравнение — Zalani, Mittal):



Спасибо за внимание!