POLITECNICO
MILANO 1863

Automatic Rhythm Transcription

Summary

Introduction to the problem of transcription

Beat Tracking and Musical Meter Analysis

Unpitched instrument transcription

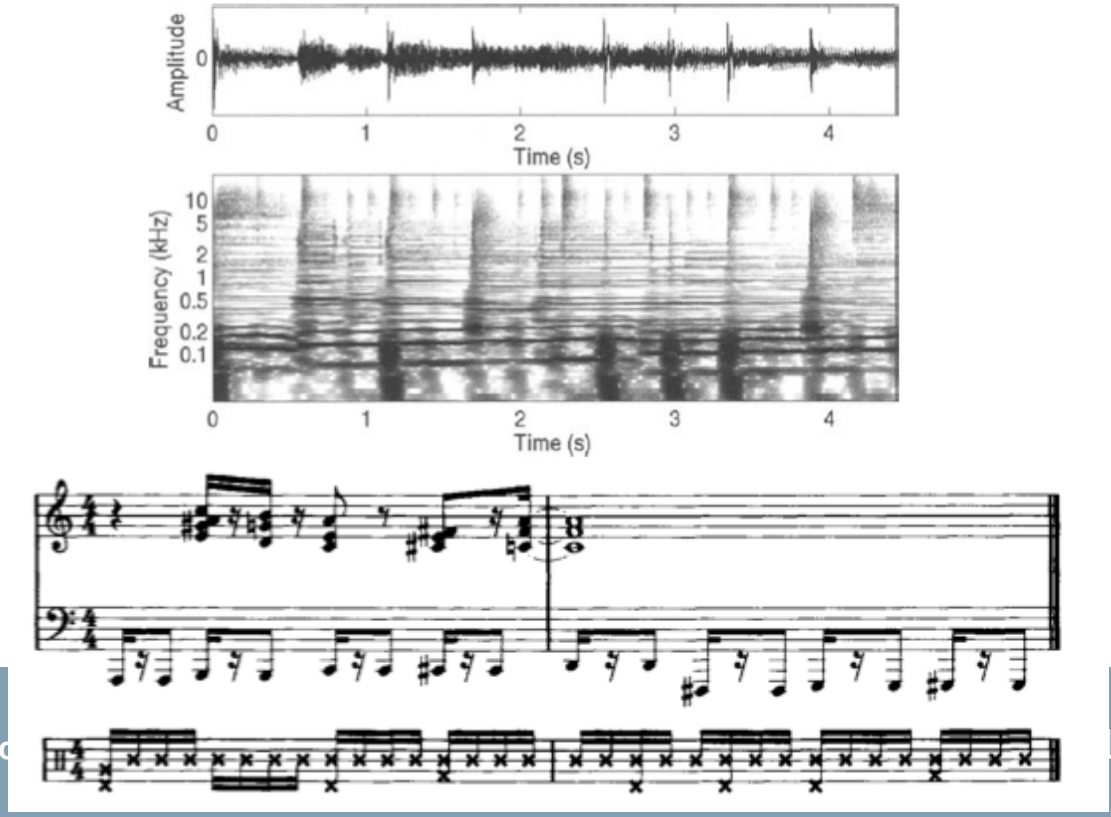
POLITECNICO
MILANO 1863

IMAGE AND SOUND
ISPG
PROCESSING GROUP

Introduction

Introduction to Music Transcription

Music transcription refers to the analysis of an acoustic musical signal in order to write down the pitch, onset time, duration, and source of each sound that occurs in it. In Western tradition, written music uses note symbols to indicate these parameters in a piece of music.



Introduction to Music Transcription

There are different types of transcription:

- Classical Notation
- Drum notation
- Guitar Tablature – Chords symbols

Common to all these representations is that they capture musically meaningful parameters

Music transcription can be seen to perform the “reverse-engineering” of a musical signal (de-mixing)

De-mixing it’s a hard task – Automatic transcription is an hard task

Two main areas of study are involved:
- Signal processing
- Music perception

Introduction to Music Transcription

For practical reasons, the scope is limited to Western music

To give a reasonable estimate of the achievable results, it is useful to study what human listeners are able to do in this task.

An average listener perceives a lot of musically relevant information in complex audio signals. He or she can tap along with the rhythm, hum the melody (more or less correctly), recognize musical instruments, and locate structural parts of the piece, such as the chorus and the verse in popular music.

Other features are for skilled listeners:

- Pitch intervals
- Chords composition
- Harmony
- Musical Style
- …

History

First attempts towards the automatic transcription of polyphonic music were made in the 1970s from Moorer. Until the end on ‘80s the number of concurrent voices was limited to two and the pitch relationships of simultaneous sounds were restricted in various ways

The first algorithm for beat tracking in general audio signals was proposed by Goto and Muraoka in the 1990s

First attempts to transcribe percussive instruments were made in the mid-80’s by Schloss

Transcription of polyphonic percussion tracks was later addressed by Goto and Muraoka

Methods

Starting for the musical signal and using different signal processing elements to extract features, many techniques are developed for automatic music annotation:

Statistical Methods

Computational models of the human auditory system

Model the human *auditory scene analysis (ASA) ability.* The term ASA refers to the way in which humans organize spectral components to their respective sounds sources and recognize simultaneously occurring sounds

More recently, several *unsupervised learning* methods have been proposed where a minimal number of prior assumptions are made about the analyzed signal

The state-of-the-art music transcription systems are still clearly inferior to skilled human musicians in accuracy and flexibility.

Some Notions

Music Transcription attempts to transcribe musical events. We need to extract some information.

- Pitch
- Loudness
- Duration
- Timbre (for instrument classification-source separation)

Musical information is generally encoded into the *relationship* between individual sound events and between larger entities composed of these.

- Bar Measure
- Pitch relationship are used to make up melodies and chords
- Inter-onset Interval (IOI) relationships
- Rhythmic pattern
- …

Some Notions

Musicological information is important for the automatic transcription of polyphonically rich musical material.

The probabilities of different notes occurring concurrently or in sequence can be straightforwardly estimated, since large databases of written music exist in an electronic format. Also, there are a lot of musical conventions concerning the arrangement of notes for a certain instrument within a given genre. In principle, these musical constructs can be modeled and learned from data.

In addition to musicological constraints, internal models may contain information about the physics of musical instruments, and heuristic rules, for example that a human musician has only ten fingers with limited dimensions.

POLITECNICO
MILANO 1863

IMAGE AND SOUND
ISPG
PROCESSING GROUP

Beat Tracking and Musical Meter Analysis

Introduction

Imagine you are sitting in a bar and your favorite song is played on the jukebox. It is possible that you might start tapping your foot in time to the music. This is the essence of beat tracking and it is a quite automatic and subconscious task for most humans.

Unfortunately, the same is not true for computers;

- replicating this process algorithmically has been an active area of research for well over twenty years

Categorizations:

- rule-based
- autocorrelative
- oscillating filters
- histogramming
- multiple agent
- probabilistic

Rhythmic Structure

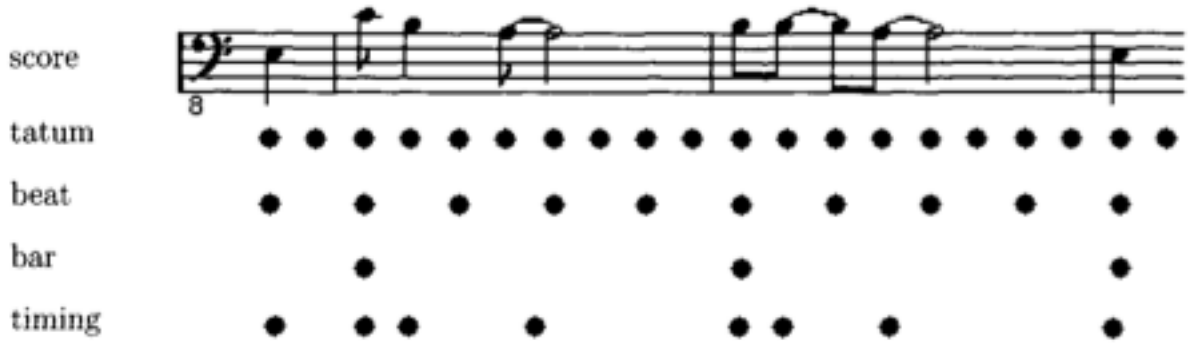
The metrical structure can be broken down into a set of three hierarchical levels

beat or ***tactus*** as the preferred (trained) human tapping tempo – basic fraction of the time signature (for a 4/4 time signature, the beat is ¼)

tatum is defined to be the shortest commonly occurring time interval

bar *or* **measure** a segment of time defined by a given number of beats of a given duration

timing is the time at which a musical event occurs



Onset detection

First we identify the onset of the notes and then we track the following set of discrete impulses.

When this approach is used, the success of any beat tracker is dependent upon the reliability of the data that is provided as input. Thus, detecting note starts in the audio can be as important as the actual beat-tracking algorithm.

Note: *harmonic* and *percussive*.

- Percussive sounds are usually characterized by significant increases in signal energy (a 'transient') and methods for detecting this type of musical sounds are relatively well developed
- Harmonic change is associated to small energy variations and therefore is much harder to be detected

Onset detection – Transient Events

Transient events, such as drum sounds or the start of notes with a significant energy change (e.g. piano, guitar), are easily detected by examining the signal envelope

An energy envelope function $E_j(n)$ is formed by summing the power of frequency components in the spectrogram for each time slice over the range required

$$\mathrm{E}_j(n) = \sum_{k \in \kappa_j} |\mathrm{STFT}_x^{\mathrm{w}}(n,k)|^2$$

Where $STFT_x^w(n,k)$ is the short-time Fourier transform (STFT) of the signal x(n) with rectangular window w centered at time n; k is the frequency index

Frame = 20 ms
Overlap = 50-75%

Use 5-10 sub-bands j that are distributed uniformly on a logarithmic frequency scale to catch different information of the transition

Onset detection – Transient Events

Uses a three-point linear regression to find $D_j(n)$, the gradient of $E_j(n)$, and peaks in this function are detected

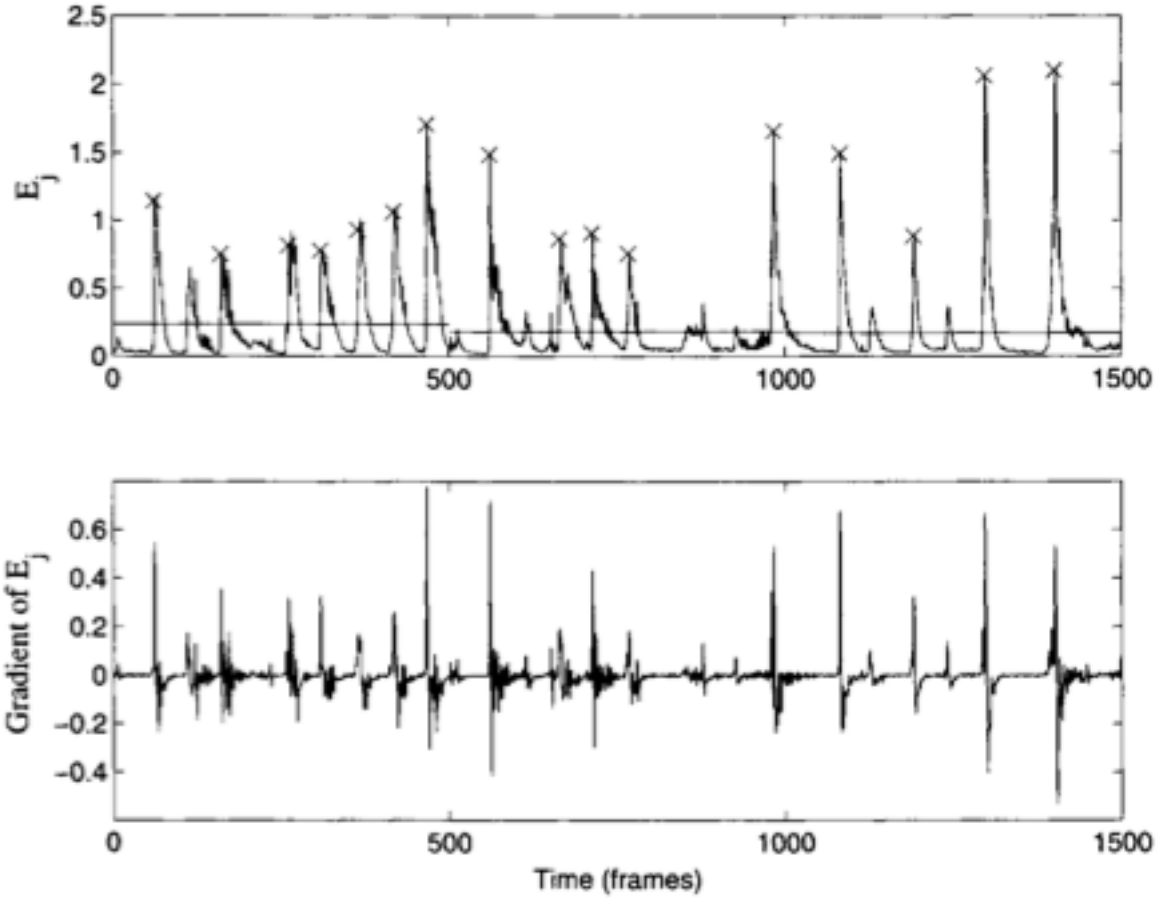
$$D_j(n) = \frac{E_j(n+1) - E_j(n-1)}{3}$$

The linear regression approach aims to detect the start of the transient, rather than the moment it reaches its peak power.

$D_j(n)$ is called a **detection function**

- Onset detection is done by simply selecting maxima in $D_j(n)$ and discarding peaks which do not pass a series of tests.
  - Threshold on energy
  - Test if there is a higher-energy peak in the neighborhood

Onset detection – Transient Events



Onset detection – Pitched Events

Note start also when there is no associated energy transient. As a consequence this tougher problem has received less attention than the detection of percussive events.

A notable recent exception is from Klapuri, who used very narrow frequency bands to detect changes in frequency

Modified Kullback-Leibler distance measure

$$\mathrm{d}_n(k) = \log_2\left(\frac{|\mathrm{STFT}_x^{\mathrm{w}}(n,k)|}{|\mathrm{STFT}_x^{\mathrm{w}}(n-1,k)|}\right)$$

$$\mathrm{d}_{\mathrm{MKL}}(n) = \sum_{k\in\mathcal{K},\mathrm{d}(k)>0} \mathrm{d}_n(k),$$

where $STFT_x^w(n,k)$ is the STFT computed within the window w. The measure emphasizes positive energy changes between successive frames and defines the spectral range over which the distance is evaluated

Peak Detection is then accomplished

Frame = 90 ms
Overlap = 87.5%

Autocorrelation method

Autocorrelation is a method for finding periodicities in data

The basic approach is to define an energy function E(n) to which local autocorrelation is then applied (in frames of length T_w, centred at time *n):*

$$\mathrm{r}(n,i) = \sum_{u=-(T_\mathrm{w}/2)+1}^{T_\mathrm{w}/2} \mathrm{E}(n+u)\mathrm{E}(n+u-i)$$

The value of i which maximizes r(n, i) should correspond to the period-length of a metrical level. This will often be the beat, but it is possible that if the tatum is strong, the autocorrelation will pick the tatum instead of the beat.

Multiple Agent

The basic philosophy is to have a number of agents or hypotheses which track independently

These maintain an expectation of the underlying beat process and are scored with their match to the data.

At the end of the signal, the agent with the highest score wins and it is chosen

An example will be provided in the next Laboratory

POLITECNICO
MILANO 1863

IMAGE AND SOUND
ISPG
PROCESSING GROUP

Unpitched instruments transcription

Pattern Recognition Approach

In general, the percussion transcription problem can be characterized by two questions:

When did something happen in the music?
Which was the event that took place?

1. Segment the input signal into events by
   1. locating potential sound event onsets in the input signal, or
   2. generating a regular temporal grid over the signal
2. Extract a set of features from each segment.
3. Classify the contents of each segment based on the extracted features.
4. Combine the segment time stamps with information about their content to yield the transcription.

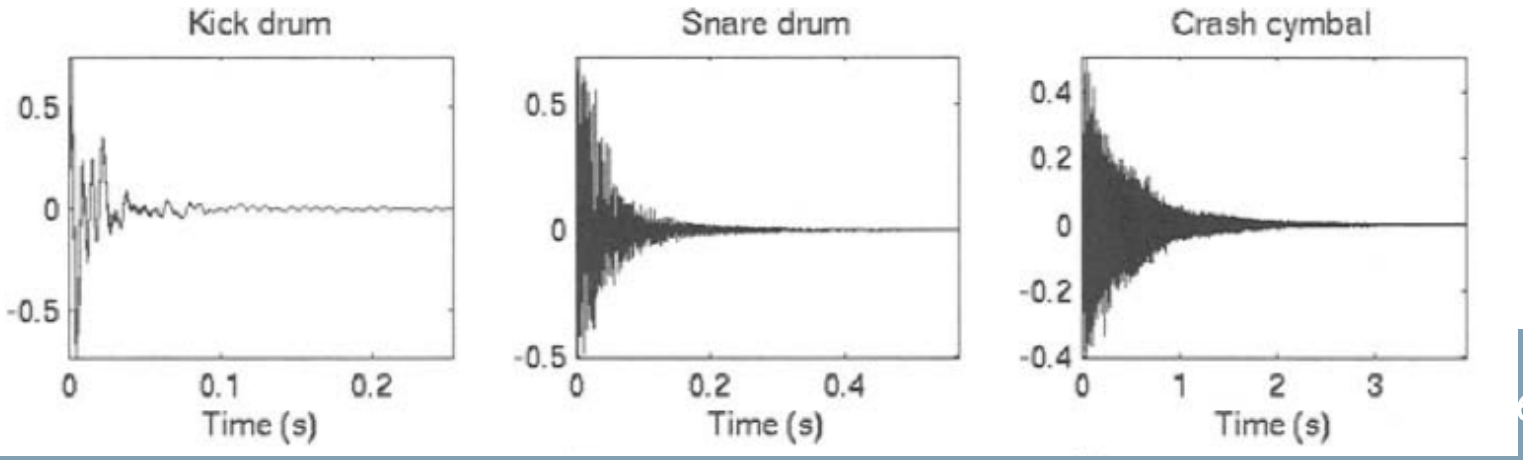
Introduction

The percussion instruments can be divided into two main types: membranophones and idiophones.

- Membranophones typically consist of a membrane or skin stretched across a frame.
- Idiophones are typically rigid bodies, such as a metal plate.

In both cases, sound is produced by striking the membrane or plate

The striking of a given drum can be modeled as an impulse function, as a broad range of frequencies will be involved in the sound.



Introduction

Approaches can be divided into two categories:

- **pattern recognition** system applied to sound events – Applied to percussion tracks
- **separation-based** system – Applied to a mixture with the presence of pitched instruments

Systems that use a **supervised approach**, through the use of trained classifiers or instrument templates, and systems that use an **unsupervised approach**, such as clustering similar segments followed by recognition of the clusters

Low-level signal analysis may not always yield a satisfying result, so some attempts have been made to utilize musicological modeling to take into account the predictability of drum patterns in music

Pattern Recognition Approach

1- Temporal Segmentation

There are two main approaches:

Onset detection - should be able to identify all the beginnings of meaningful sound events and still be robust against noise which might generate extraneous on- sets – Described earlier

Generate a temporal grid over the whole signal and use it for the segmentation

- The grid spacing is determined by the fastest rhythmic pulse present in the signal, so that almost all the events in the piece coincide with a grid point – **Tatum grid**
- Only events that occur on the tatum grid, or in the proximity of the time instants of the grid, are considered.

Pattern Recognition Approach

There are several ways for estimating the **Tatum** pulse

- Detect onsets in the input signal
- Calculate time intervals between all onset pairs
- Determine the period of the Tatum as the greatest common divisor of the time intervals.
- The phase of the grid is then estimated by aligning it with the located onsets.

- Another method of Tatum estimation uses a bank of comb filter resonators and probabilistic modeling to find the Tatum period and phase

Pattern Recognition Approach

Pros & Cons of Tatum Grid

Overestimating the Tatum period causes severe errors in the segmentation. Another drawback is that expressive playing causes deviations from the equidistant grid

An advantage of the grid representation is that it is less prone to errors caused by inserting and deleting sound onsets than the onset detection-based approach: events wrongly recognised by the onset detection are discarded since they are not aligned on the tatum grid.

Segmentation

Meaningful parts of the signal need to be segmented

The simplest approach is to take a part of the signal starting at the located onset or grid point, and ending at the next located onset or grid point

Due to the impulsive nature of unpitched sounds, windowing techniques are generally not used

Pattern Recognition Approach

2 - Feature Extraction

The aim of feature extraction is to obtain numerical values describing the segments so that they can be recognized or grouped together, while reducing the amount of irrelevant information in the time-domain signal

Frequency domain features

Mel-frequency cepstral coefficients (MFCCs) describe the rough shape of the signal spectrum

bandwise energy descriptors. The energy content of the sound is calculated in a few frequency bands and their relations to the total signal energy are used as features. The number of bands and their spacing greatly depends on the desired frequency resolution

spectral centroid, spectral spread, spectral skewness

Pattern Recognition Approach

Time domain features

Temporal Centroid - describes the temporal balancing point of the sound event energy

$$C_t = \frac{\sum_t tE(t)}{\sum_t E(t)}$$

where E(t) denotes the root-mean-square (RMS) level of the signal in a frame at time t. It enables discrimination between short, transient-like sounds and longer ringing sounds

Zero Crossing Rate - It correlates with the spectral centroid and the perceived brightness of the signal (Snare – Bass Drum)

It was noticed that in most cases, using a feature set that has been chosen through some feature selection method yielded better results than using all the available features. Also, a **dimension reduction** method such as principal component analysis can be applied to the set of extracted features prior to classification

Pattern Recognition Approach

3- Segment Recognition

The extracted features are then used to recognize the percussive sounds in each segment

Two different ways:

- detecting the presence of a given drum, even if other drums occur at the same time (instruments separation)
- recognizing drum combinations directly

Second approach could be easier but the number of possible combinations could be very large

Three possible classification categories:

- decision tree methods
- instance-based methods
- statistical modeling methods

Pattern Recognition Approach

Decision tree Methods

operate by asking a sequence of questions about the sample to be classified - *Is the spectral centroid of the signal above 500 Hz?*

Each answer rules out some of the possible classification results and defines the next question

Instance-based Methods

Given a training set, it determines the label for the analysed sample by comparing it to the training data - SVM

Statistical Modelling Methods

Use properties of statistical distribution of features

GMM - model the distribution of the feature values in each class as a sum of Gaussian distributions

Pattern Recognition Approach

Statistical Modeling Methods

GMMs are used in conjunction with hidden Markov models (HMM) to represent the feature distributions in their states. A basic HMM consists of two parallel processes: a hidden state process which is assumed to be a Markov chain and cannot be directly observed, and an observation process (features). The observed features are conditioned on the hidden state by using a GMM in each state and, based on the observations, the hidden state sequence can be inferred

Template Matching Method

Given a Mixture of percussive instruments

Given the templates for each drum type obtained from a power spectrogram of each drum type in isolation

A distance measure is used to detect the presence of the template in the mixture

Separation Based Approach

Separation-based techniques aim to separate the instruments into distinct streams through the analysis of mixtures of drums or percussive instruments.

It is important to note that separation in the context of transcription means the separation of frequency and amplitude characteristics associated with each source in order to identify and transcribe them

Here we are not interested in de-mixing and re-synthesis, but these methods could also be used for those purposes

All these techniques assume that the mixture spectrogram matrix X of size (KxT), where K is the number of frequency bins and T is the number of time frames, results from the superposition of J source spectrograms Y_j of the same size

We can define X as $\mathbf{X} = \sum_{j=1}^{J} \mathbf{Y}_j = \sum_{j=1}^{J} \mathbf{b}_j \mathbf{g}_j^T$

- b_j is an invariant frequency basis function – length(b_j) = K
- g_j are invariant time basis functions – length(g_j) = T

Separation Based Approach

The techniques differ in how the decomposition of the spectrogram X is accomplished

Use of invariant frequency basis functions → no pitch changes are allowed over the course of individual spectrograms Y_j → drum sounds

this type of decomposition particularly suits to analyse percussive tracks in polyphonic music.

Many methods:

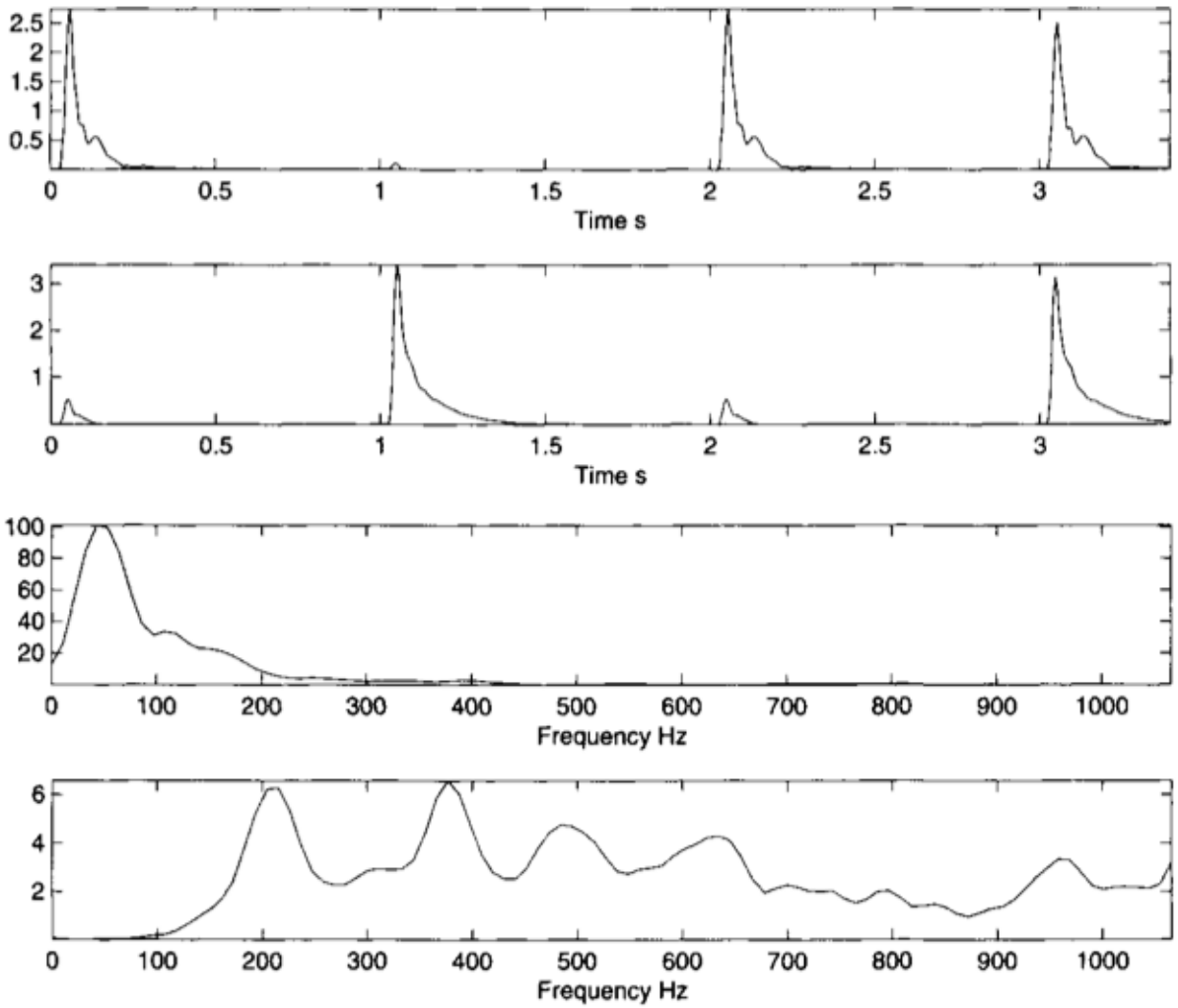
ISA –Independent Sub-space analysis

- PCA (Principal Component Analysis) on the spectrogram, keeping only a small number of frequency lines, and then performing ICA (Independent Component Analysis) on the retained components

NMF – Non negative matrix factorization

- Nmf(X) = HW – Factorization of a Matrix into two matrices

Separation Based Approach



Kick g_1

Snare g_2

Kick b_1

Snare b_2

Separation Based Approach

Prior Subspace Analysis (PSA)

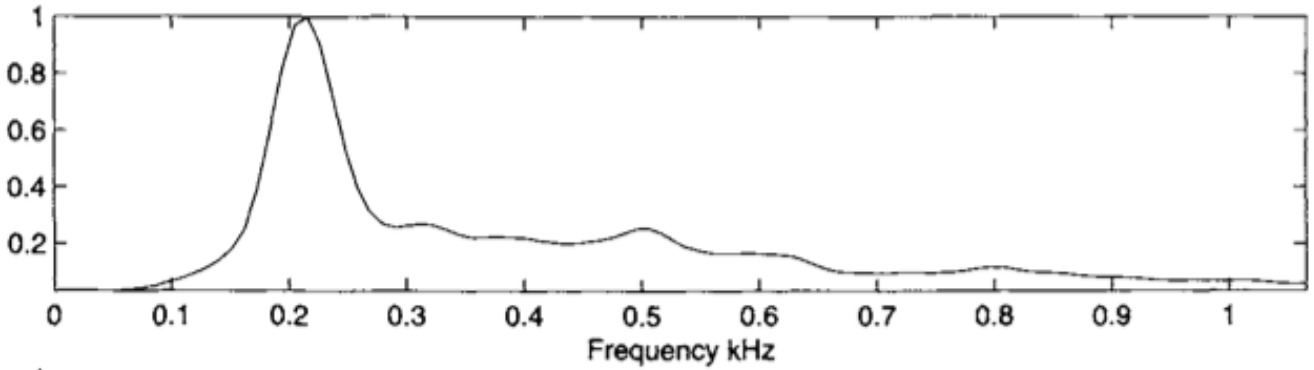
PSA assume that a priori knowledge b_{prj} of b_j exists

$$\mathbf{X} \approx \sum_{j=1}^{J} \mathbf{b}_{\mathbf{pr}j}\mathbf{g}_j^{\mathsf{T}}$$

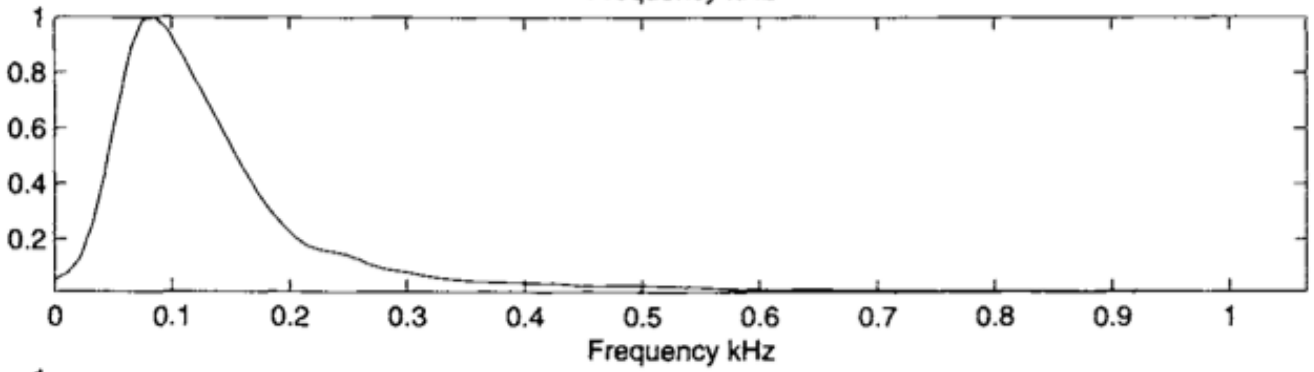
These priors are obtained by performing ISA on a large number of isolated samples of each drum type and retaining the first frequency basis functions from each sample. The priors shown then represent the average of all the frequency basis functions obtained for a given drum type

We can use other matrix-factorization techniques such as NMF

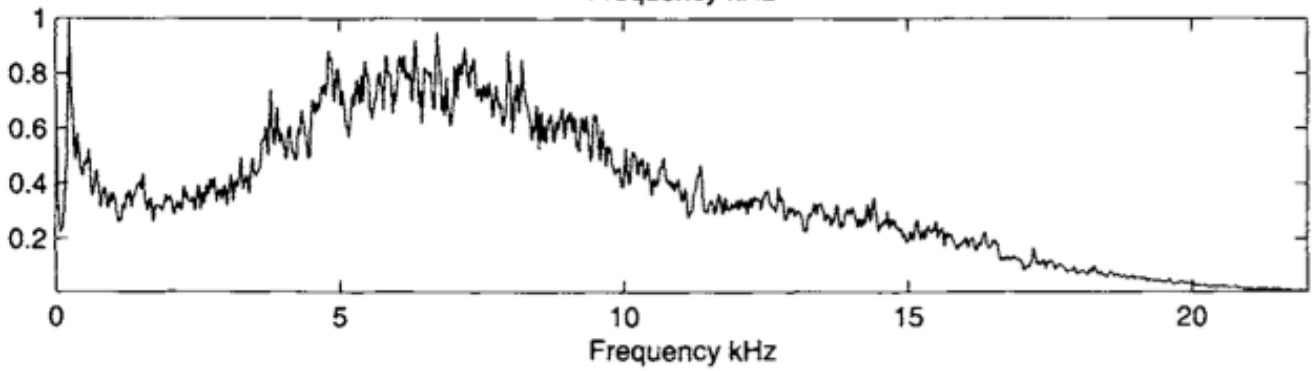
Separation Based Approach



Snare



Kick



Hi Hat

Separation Based Approach

PSA attempts to estimate the amplitude basis functions

$$\mathbf{X} \approx \mathbf{B_{pr}}\mathbf{G},$$

$$\mathbf{B_{pr}} = [\mathbf{b_{pr1}}, \ldots, \mathbf{b_{prJ}}]$$
$$\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_J]^T$$

Estimates of the amplitude basis functions

$$\hat{\mathbf{G}} = \mathbf{B_{pr}}^{+}\mathbf{X},$$

Where B_{pr}^{+} is the pseudoinverse of B obtained by $B^{+}=(B^{T}B)^{-1}B^{T}$ - assuming that b_{prj} are linearly independent

Given the broad-band nature of drum sounds, the occurrence of a given drum will cause a partial match with the prior subspace of another drum $\hat{g}_j$ the estimated amplitude basis functions → are not independent

Separation Based Approach

To overcome this problem ICA is carried out on the estimated amplitude basis functions $\hat{G}$

$$\mathbf{G} = \mathbf{W}\hat{\mathbf{G}},$$

Where W is the de-mixing matrix obtained from ICA

Now an adaptive process is used to improve the frequency basis functions

$$\mathbf{B} = \mathbf{X}\mathbf{G}^{+}.$$

Prior subspaces are generated and PSA is performed on the input signals. Once good estimates of the amplitude basis functions had been recovered, onset detection is carried out on these envelopes to determine when each drum type is played.

Separation Based Approach

Prior Subspace Analysis (PSA) in presence of pitched instruments

The presence of a large number of pitched instruments will cause a partial match with the prior subspace used to identify a given drum

It should be noted that pitched instruments have harmonic spectra with regions of low intensity between the overtones or partials

good frequency resolution reduces the interference due to the pitched instruments

In Bass Drum and Snare Drum transcription setting all values in the initial estimates of the amplitude basis functions $\hat{G}$ below a fixed threshold are set to zero

Harder for hi-hat and cymbals

Musicological Modeling

Music events are connected in structure through some relationships

Introducing musicological information can improve the transcription process

Use the concept of Pattern

- Use classification systems to classify possible patterns
- Use Markov Chains as in speech recognition methods – Given a drum event Markov Chains defines likelihoods for next possible events