POLITECNICO
MILANO 1863

Automatic Chord Estimation

Summary

- Definition of chords
- Problem formulation
- Feature extraction: chromagram and its variants
- Chord estimation: Hidden Markov Models and their variants.

POLITECNICO
MILANO 1863

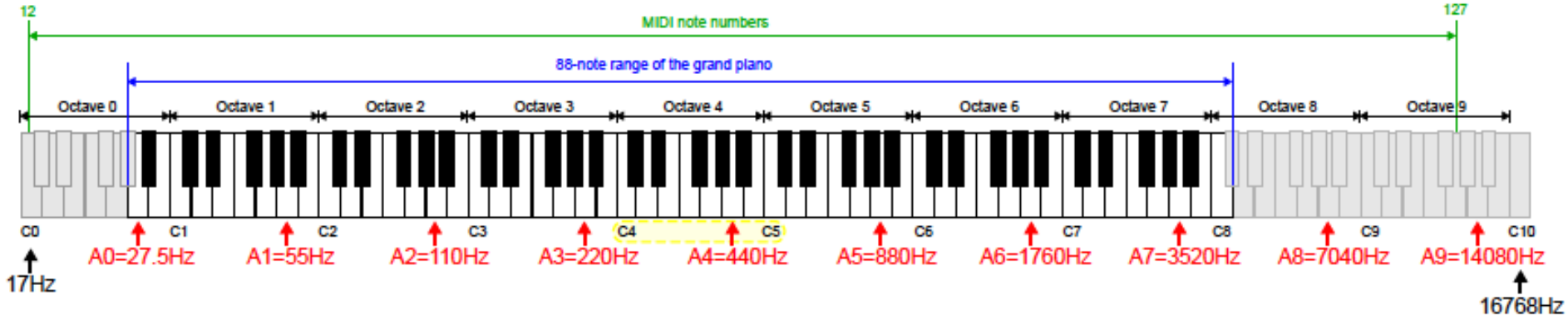
IMAGE AND SOUND
ISPG
PROCESSING GROUP

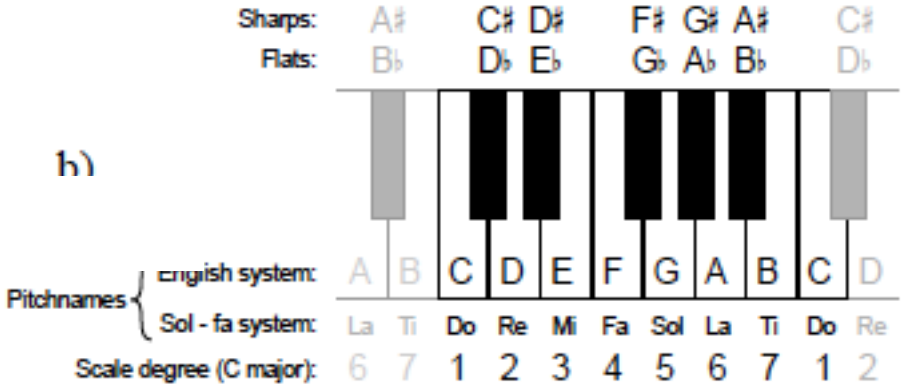
Introduction

Notes and musical scales

- Note: the combination of a pitch determining the fundamental frequency of the note and a duration that determines the length of time that the pitch is sounded for.
- A pitch is defined by a pitch name and an octave number. A pitch name comprises a natural name plus zero or more sharps (♯) or flats (♭).

Notes and musical scales





From [Harte2010]

Notes and musical scales

Smallest difference between two pitches on a piano keyboard: semitone.

Relative difference between two pitches: interval.

Degree		Name	Relation to Tonic
1 First	I	Tonic	Unison
2 Major second	II	Supertonic	One tone above the tonic
(♭3 Minor third	♭III)	Mediant	Mid way between tonic and dominant
3 Major third	III	"	" " " " " "
4 Perfect fourth	IV	Subdominant	Fifth below the tonic
5 Perfect fifth	V	Dominant	Fifth above the tonic
(♭6 Minor sixth	♭VI)	Submediant	Mid way between the subdominant and the tonic
6 Major sixth	VI	"	" " " " " " " " "
(♭7 Lowered Seventh	♭VII)	Subtonic	One tone below the tonic
7 Seventh	VII	Leading Note	Leads into the tonic

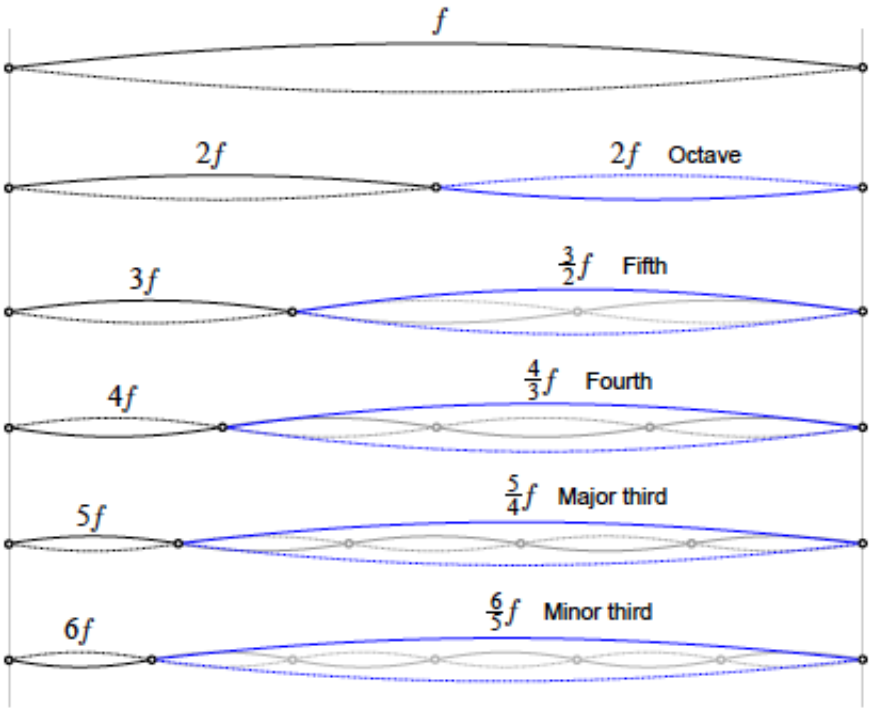
Notes and musical scales

- The perfect fourth and fifth intervals are made with the tones which are closest to the prime (the tonic reference note for the interval) in the harmonic series.
- The Major (M) intervals are intervals formed by the tonic and the notes of the major scale other than the fourth and fifth.
- The Minor intervals (m) are one semitone less than their corresponding major interval.

Notes and musical scales

When a string is plucked, the resulting pitch that we hear is not just the fundamental frequency, but a number of different frequency components.

The string will vibrate at frequencies that are integer multiples of the fundamental frequency which are known as harmonics or harmonic overtones or partials.



Chord definition

Definition of chord given by the Oxford Dictionary of Music :

Any simultaneous combination of notes, but not usually fewer than three. The use of chords is the basic foundation of harmony

Consonance and dissonance

The degree to which a simultaneous combination of notes is perceived to be acceptable or pleasing in a given musical context is called consonance and its converse i.e. how unpleasant it is, is dissonance.

Consonance and dissonance

In psychoacoustic experiments during the 1960s, Plomp and Levelt linked the perceived consonance of two simultaneously sounding sine waves to the critical bandwidth in human hearing:

- Sounds with a frequency separation greater than the critical bandwidth were judged as consonant
- Sounds with the same frequency are judged as perfectly consonant
- Sounds with a frequency separation between 5% and 50% of the critical bandwidth extension are judged as dissonant.

Consonance and dissonance

Musical instruments produce complex sounds containing many harmonically related frequency components.

For a combination of notes, all of the audible harmonics (up to the sixth or seventh) of each note contribute to the perceived consonance or dissonance of the chord and each pair of harmonic components must adhere to the rules determined above.

Chords

- The two most common types of chord are the major and the minor triads.
- The major triad is made up of the tonic note plus the fifth and third degrees of the major scale; in the key of C the tonic major triad is CEG.
  - This triad can be viewed as a major third interval (I-III) underneath a minor third (III-V).
- The minor triad is made up of the tonic note plus the fifth and the third of the minor scale which, with root note C is CE ♭ G.
  - This triad is the complement of the major triad in that it is built from a minor third underneath a major third

Chords

An augmented chord is built up of major thirds; the C augmented chord would thus contain the notes CEG♯.

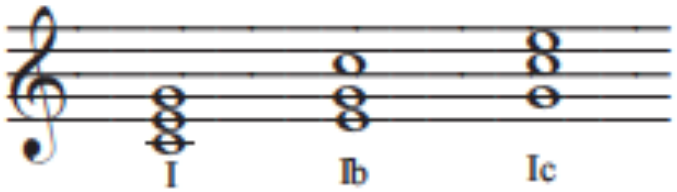
The diminished chord is built up of minor thirds, hence the C diminished triad would be CE ♭ G ♭ .

The suspended second chord is a major second under a perfect fourth e.g. CDG.

A suspended fourth is a perfect fourth underneath a major second e.g. CFG

Chords

Inversions: the bass note in the chord does not have to be necessarily the root of the chord: any of the chord notes can be used as the bass note



- I : the C major triad in root position
- Ib: first inversion, the third constitutes the bass note
- Ic: second inversion, the fifth constitutes the bass note

Notation: often the inversion is written as the chord name and the bass note to use. As an example, Ib can be written as C/E (spelled as "C over E")

Chords

Extension: more complex combinations of sounds can be obtained by adding more notes.

- Major 7th chord: a major 7th is added to a major triad (maj7).
- Dominant 7th chord: a minor 7th is added to a major triad (7).
- Minor 7th: a minor 7th is added to a minor triad position (min 7).

POLITECNICO
MILANO 1863

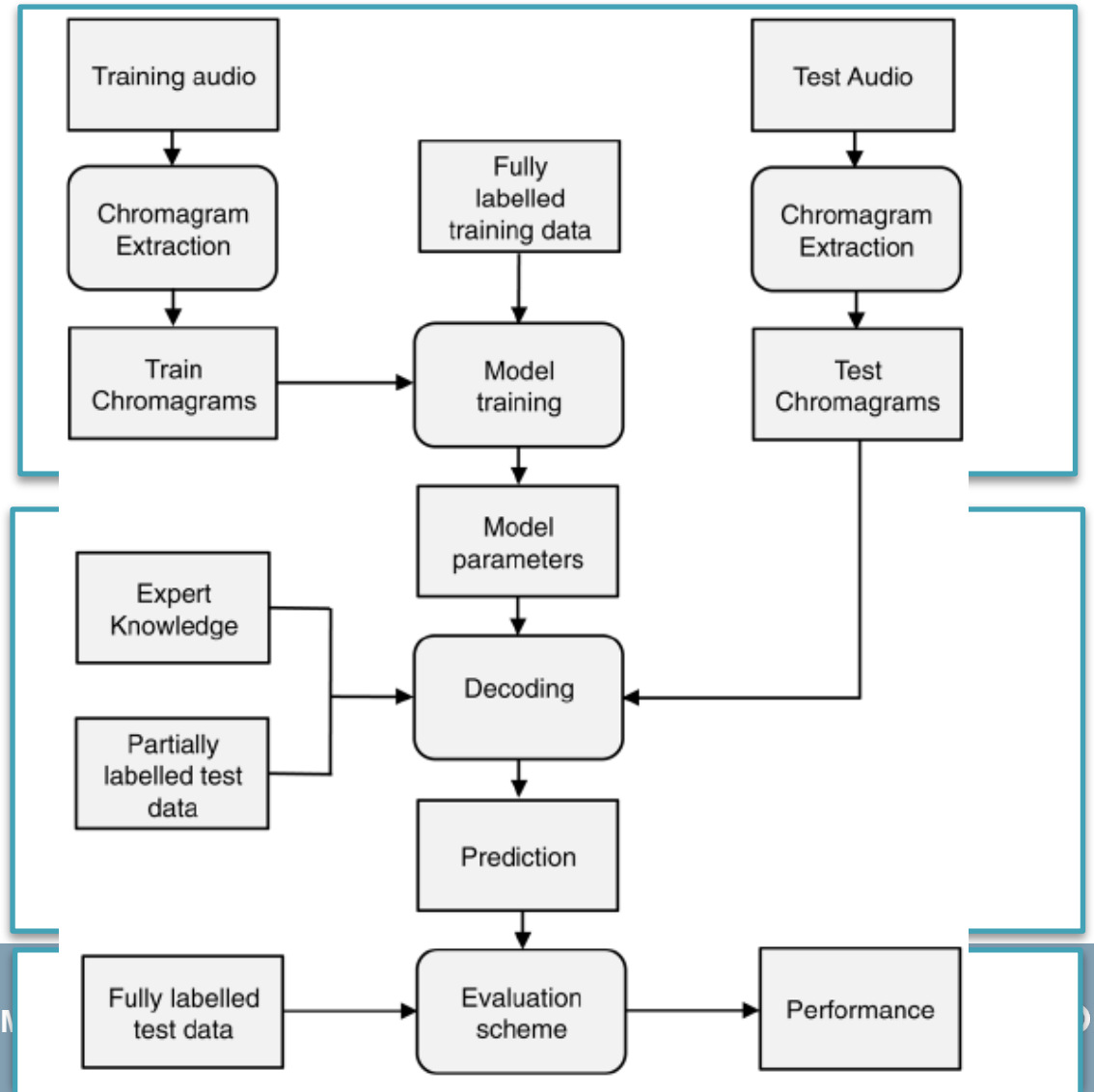
PROBLEM FORMULATION

Automatic Chord Extraction

Goal: annotate chords from a musical content in an automated fashion using machine learning techniques.

Flow diagram for Automatic Chord Extraction

From [McVicar2014]



Training

Test

Performance evaluation

Feature extraction

The feature universally used in Automatic Chord Extraction is the chromagram, and its variants.

- The chromagram describes how the pitch saliences vary across the duration of the audio.
- The chromagram can be represented with a real-valued matrix **X** containing a row for each pitch class considered, and many columns as many frames considered.
- A column **x** extracted from **X** is referred to as *chroma vector* or *chroma feature*

Modelling strategies

Goal of the modelling is to assign a label to chromagram frames.

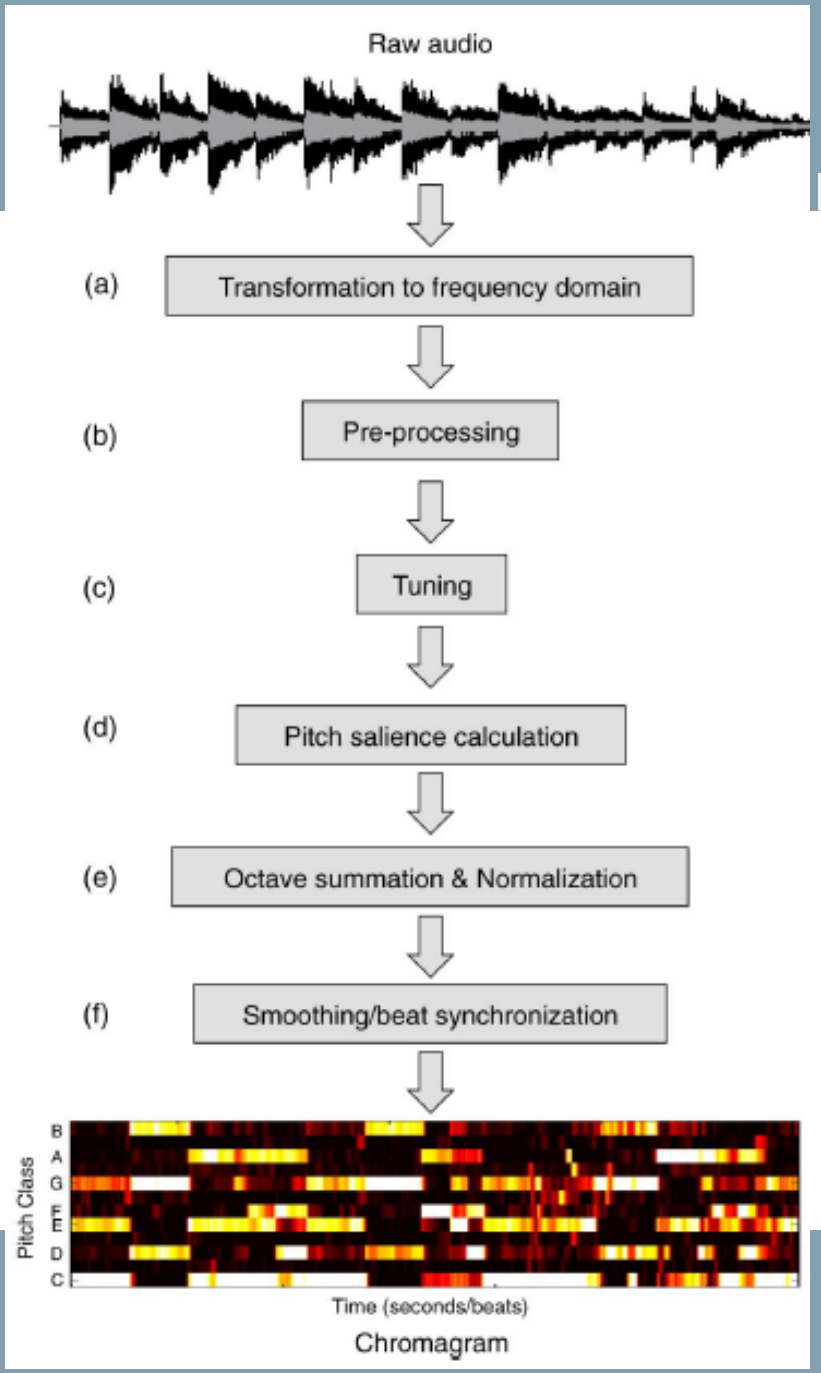
Different strategies have been adopted: template matching, Hidden Markov Models and Dynamic bayesian networks are the most popular ones.

POLITECNICO
MILANO 1863

FEATURE EXTRACTION

Chromagram

Computation scheme of the chromagram:



Transformation to the frequency domain

The most common transformation to the frequency domain of the raw signal is represented by the Short Time Fourier Transform.

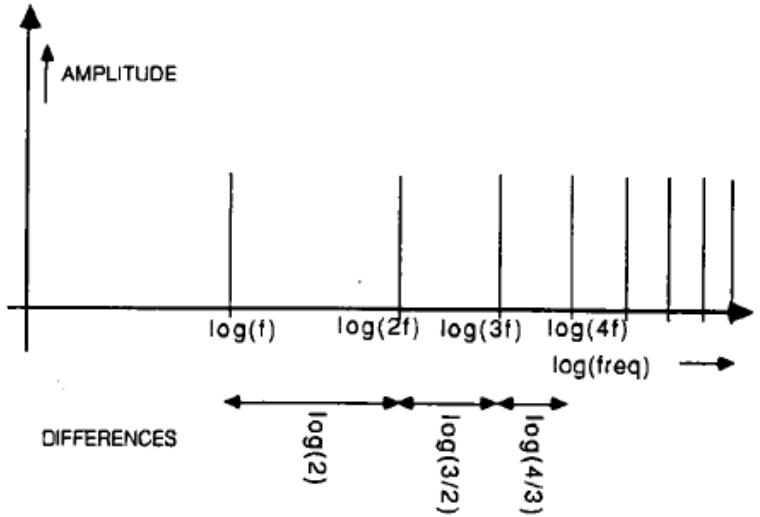
- Tradeoff between frequency and time resolution: with short windows low frequencies cannot be distinguished, while with long windows the time resolution becomes poor.
- For ACE purposes, tones that are half-a-tone apart must be distinguishable, therefore this poses a limit on the time resolution. This limit is particularly poor if we aim at capturing low frequencies.

Transformation to the frequency domain: Constant-Q spectrum

A solution that partially overcomes the problem of the STFT is the Constant-Q spectrum [Brown1991].

Spectrum of an harmonic sound on a log frequency axis:

The spacing between the fundamental and the first harmonic is log(2), the spacing between the second and the third is log (3/2), the spacing between the third and the fourth is log(4/3) and so on.
This spacing is constant whatever is the fundamental frequency.



Constant-Q spectrum

Consider a window of 1024 samples at 32 kHz: using the DFT the resolution is 31.3 Hz (32000/1024 Hz).

- At the low end of the range of a violin, the G3 is 196Hz, so the resolution is 16% of the frequency, which is much greater than the 6% of the frequency separation for two adjacent notes tuned in equal temperament.
- At the upper end of the piano range, the frequency C8 is 4186Hz, and 31.3 Hz represents the 0.7% of that frequency: far more samples are computed at this frequency range.

Goal of the Constant-Q spectrum: the resolution should be geometrically related to the frequency, e.g. 3% of the frequency, in order to distinguish between frequencies with semitone spacing.

Constant-Q Spectrum

The frequencies computed by the DFT should be exponentially spaced. If we require a quarter tone spacing, this poses a limit of (2exp(1/24) -1) = 0.029 times the central frequency.

We should impose that $f/\delta f = Q = \text{constant}$. In our case $\delta f = 0.029\, f$, therefore $Q = f/0.029f = 34$ to have a 1/24 oct. filter bank.

Constant-Q spectrum

The frequency of the *k*th spectral component is $f_k = 2^{k/24} f_{\min}$

The length of the window in samples at frequency f_k is $N(k) = S/\delta f_k = (S/f_k)Q$, where *S* is the sample rate.

Discrete FT to obtain the *k*th component of the Constant-Q spectrum:

$$X(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} W(k,n)x(n)e^{-j2\pi Qn/N(k)}$$

$$W(k,n) = \alpha + (1-\alpha)\cos(2\pi n/N(k)), \alpha = 25/46$$

Preprocessing

In a polyphonic musical excerpt, not all the components of the signal are beneficial for the understanding of harmony.

Removing the background spectrum may clean the chromagram and improve the accuracy of chord estimation and includes the following subtasks:

- Removing percussive elements of the music [Ono2008, Reed2009], also adopting Harmonic Percussive Source Separation (HPSS), modelling the signal as Harmonic plus Percussive components.
- Harmonics and subharmonics can easily confuse the chord recognition. Strategies for detecting and removing harmonics are presented in [Pauws2004], [Lee2006], [Papadopoulos2007], [Mauch2010].
- A joint removal of harmonics and percussive components is presented in [Varewyck2008].

Tuning

In order to compensate for tunings different from A4 = 440 Hz, it is possible to compute a spectrogram at twice the required frequency resolution (quarter of tone), or having at least 3 frequency bands per semitone.

Interpolation is used to infer the exact pitch salience.

Pitch salience

Even if the spectrogram provides a good representation of the pitch evolution, tentatives of mapping the spectrogram to information more closely related to the human perception have been made.

- In [Pauws2004] authors weigh the spectrogram amplitude by an arc-tangent function;
- An A-weighting in used in [Ni2012] with a great improvement in Automatic Chord Estimation.

Octave summation and normalisation

In order to work with a 12-dimensional representation of the pitch evolution of the audio, all pitch saliences belonging to the same pitch class must be summed together, disregarding the octave information.

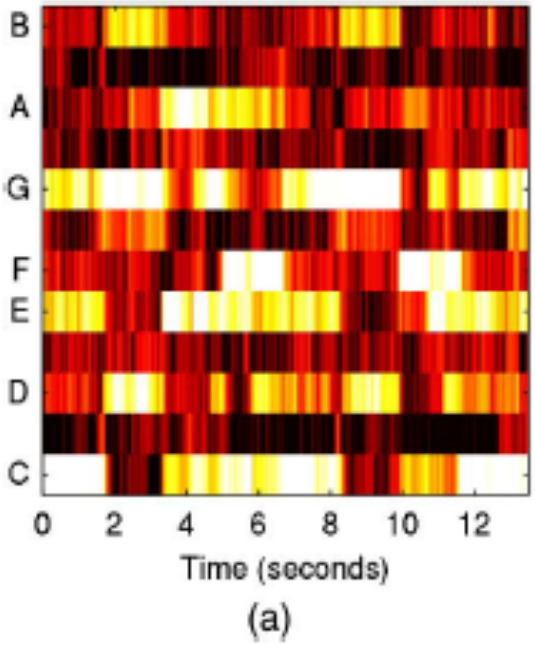
A normalization is then adopted on a frame basis (i.e. on **x**) to obtain results independent of the volume in the track.

Smoothing and beat synchronisation

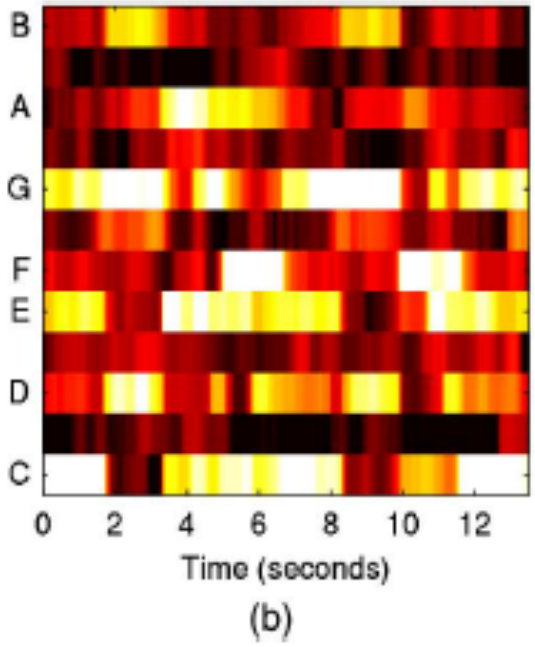
It has been noticed that the use of instantaneous chroma lead to erroneous chord changes.

- In [Fujishima1999] a smoothing of the chroma vectors has been proposed.
  - Typical operators for the smoothing are the mean and the median of the pitch salience.
  - In [Bello2011] the recurrence plots are used to find the number of times the pitch salience visits the same location in a given interval, and to perform smoothing based on this observation.
- In [Bello2005] the authors imposed that chords are beat-synchronous.

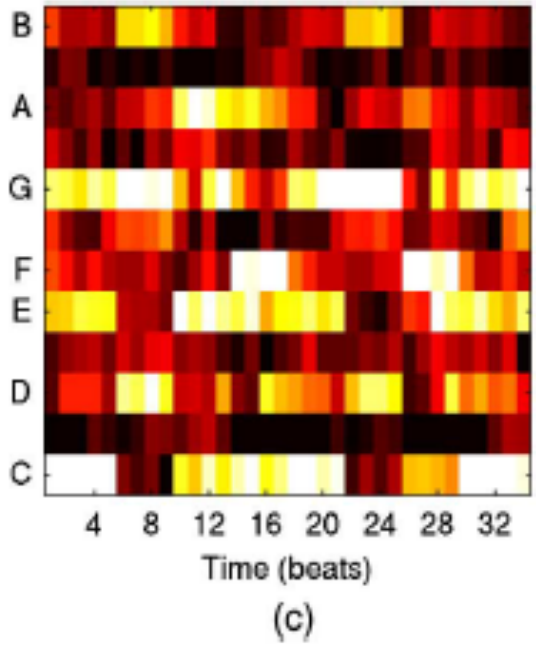
Chromagram examples



Standard chromagram



Median-based chromagram



Beat synchronised chromagram

POLITECNICO
MILANO 1863

CHORD MODELING

Goals and methodologies

Goal: assign chord to chromagram frames.

Different solutions proposed:

- Template matching
- Hidden Markov Models
- Dynamic Bayesian Networks
- Recurrent Neural Networks

Template matching

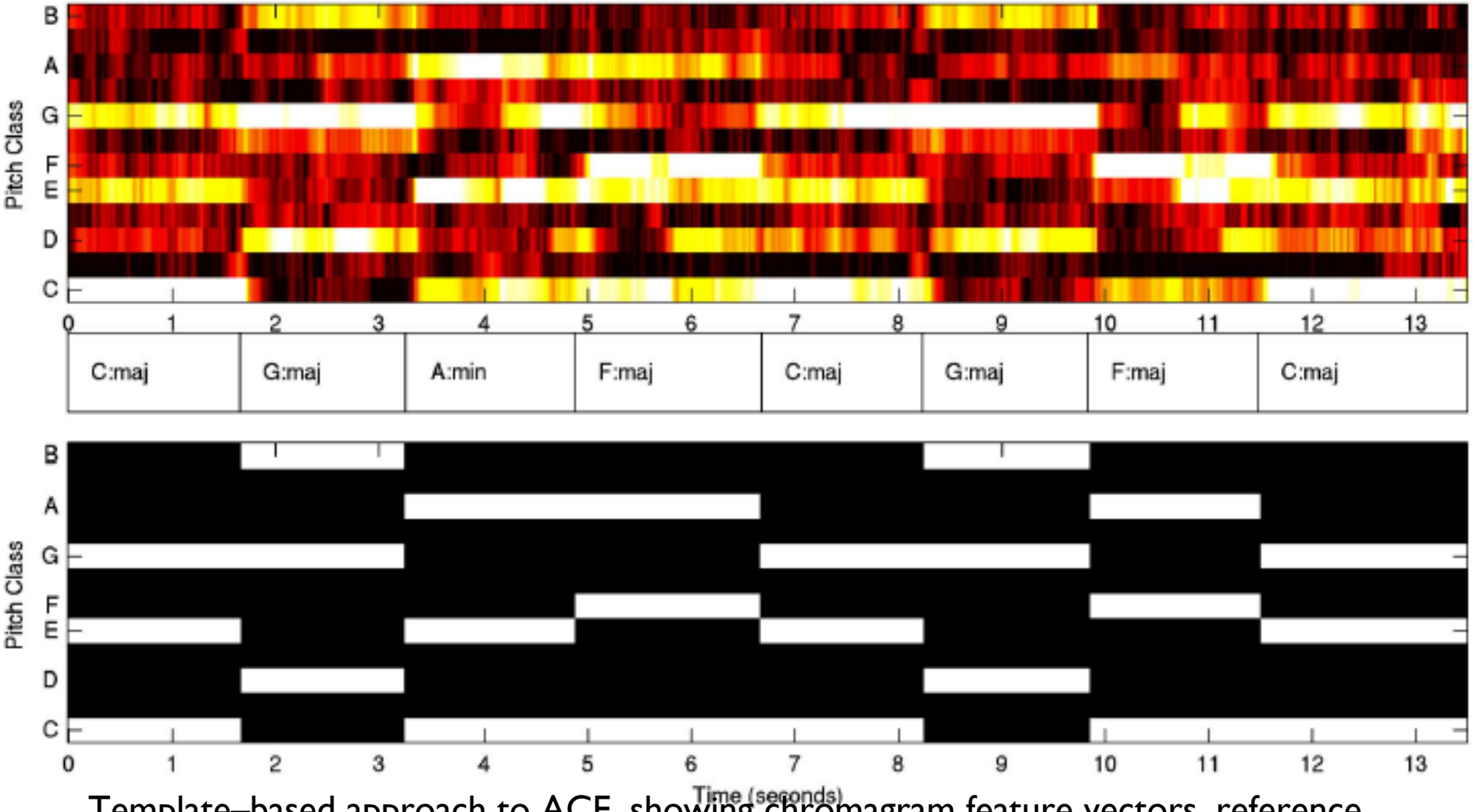
The most intuitive solution for chord recognition.

Template matching involves comparing feature vectors against the known distribution of notes in a chord.

A 12–dimensional chroma vector is compared to a binary vector containing ones where a trial chord has notes present. For example, the template for a C Major chord would be [1 0 0 0 1 0 0 1 0 0 0 0].

Each frame of the chromagram is compared to a set of templates, and the template with maximal similarity to the chroma is output as the label for this frame

Template matching: example



Template–based approach to ACE, showing chromagram feature vectors, reference chord annotation and bit mask of optimal chord templates.

Hidden Markov Models

Individual pattern matching techniques such as template matching fail to model the continuous nature of chord sequences.

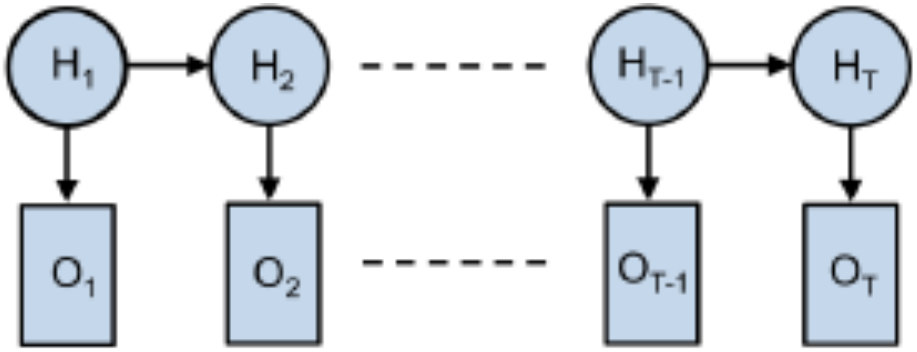
One of the most common ways of incorporating smoothness in the model is to use a Hidden Markov Model (HMM).

An HMM is a probabilistic model for a sequence of observed variables, called the observed variables.

It is assumed that there is a sequence of hidden variables, paired with the observed variables, and that each observed variable is independent of all others when conditioned on its corresponding hidden variable.

Hidden variables form a Markov chain of order 1.

Hidden Markov Models



Hi: hidden variables (chords)

Oi: observations (chromagram vectors)

Sequence of chord symbols: $\mathbf{y}$

HMMs formalize the probability distribution $P(\mathbf{y}, \mathbf{X}|\Theta)$

Hidden Markov Models

Chords are modelled as a first-order Markov chain, i.e. the future chord is dependent on the current chord only and not on the past ones.

HMMs also assume that the observation vector (i.e. the chromagram) is independent on the the other model parameters.

$$P(\mathbf{y}, \mathbf{X}|\Theta) = P_{\text{ini}}(y_1)P_{\text{obs}}(\mathbf{x}_1|y_1)\prod_t P_{\text{tr}}(y_t|y_{t-1})P_{\text{obs}}(\mathbf{x}_t|y_t)$$

$P_{\text{ini}}(y_1)$: probability that the first chord is y_1

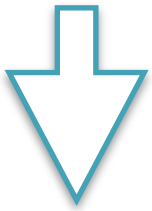
$P_{\text{tr}}(y_t|y_{t-1})$: probability that the chord is y_{t-1} is followed by y_t

$P_{\text{obs}}(\mathbf{x}_t|y_t)$: probability density for the chromagram $\mathbf{x}_t$ given the chord y_t, known as emission probabilities.

Hidden Markov Models

Assumption: the HMM is stationary

Emission probabilities: Gaussian distribution with mean vector **μ** and covariance matrix **Σ**.



HMM model parameters: $\Theta = \{\mathbf{P}_{\text{tr}}, \mathbf{P}_{\text{ini}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

$\mathbf{P}_{\text{tr}} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$ $\quad \boldsymbol{\mu} \in \mathbb{R}^{12 \times |\mathcal{A}|}$

$\mathbf{P}_{\text{ini}} \in \mathbb{R}^{|\mathcal{A}|}$ $\quad \boldsymbol{\Sigma} \in \mathbb{R}^{12 \times 12 \times |\mathcal{A}|}$

HMM training and decoding

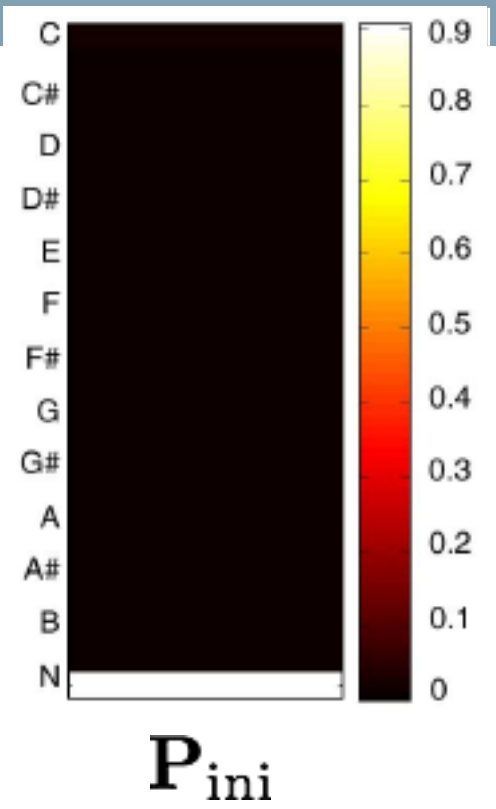
Training of the HMM corresponds to finding the parameters of the HMM given a dataset of labelled chromagrams.

- Typical solution: Baum-Welch algorithm [Baum1970,Welch2003].

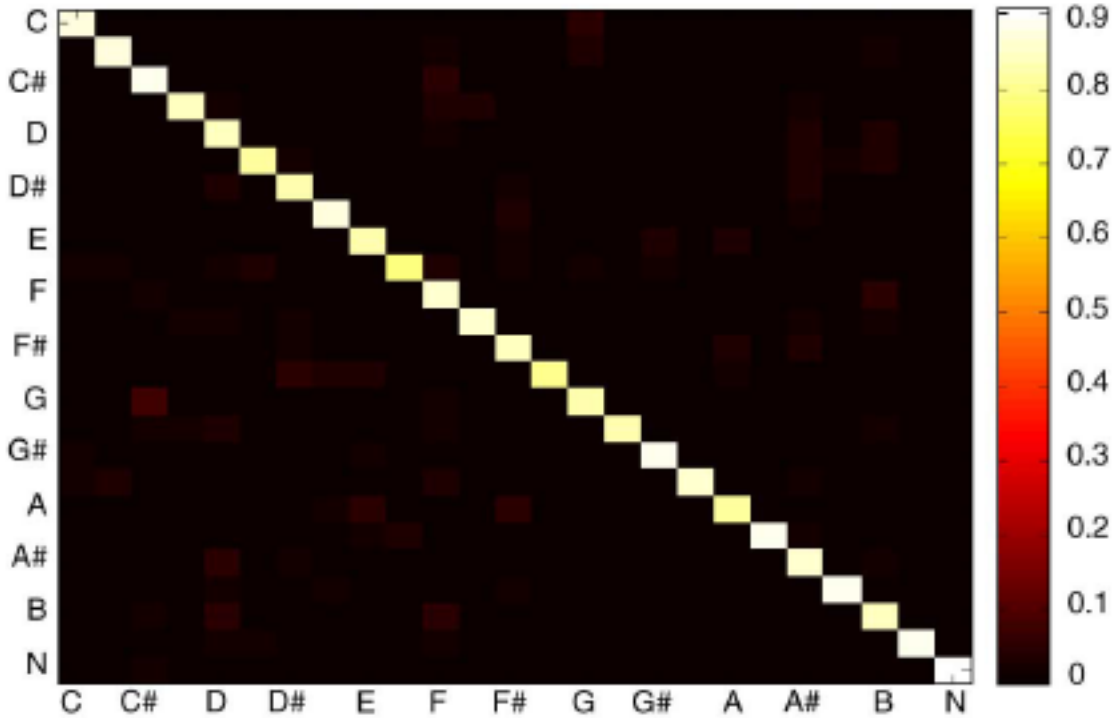
Decoding of the HMM: finding the sequence of hidden variables (chords) that best explains the observations.

- Typical solution: Viterbi algorithm [Viterbi1967]

HMM training and modelling

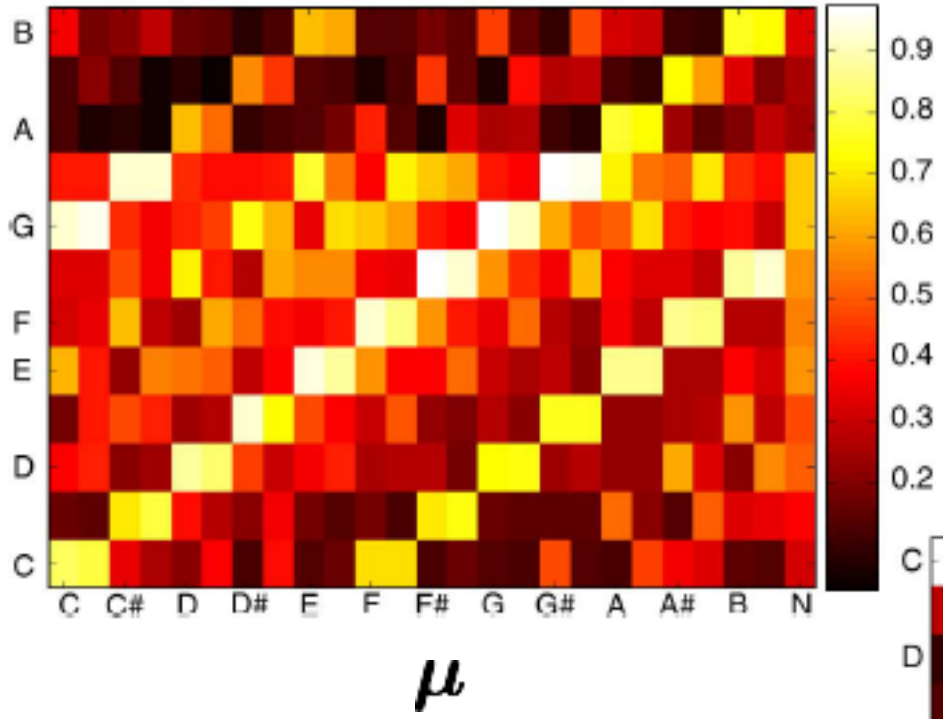


$\mathbf{P}_{\text{ini}}$



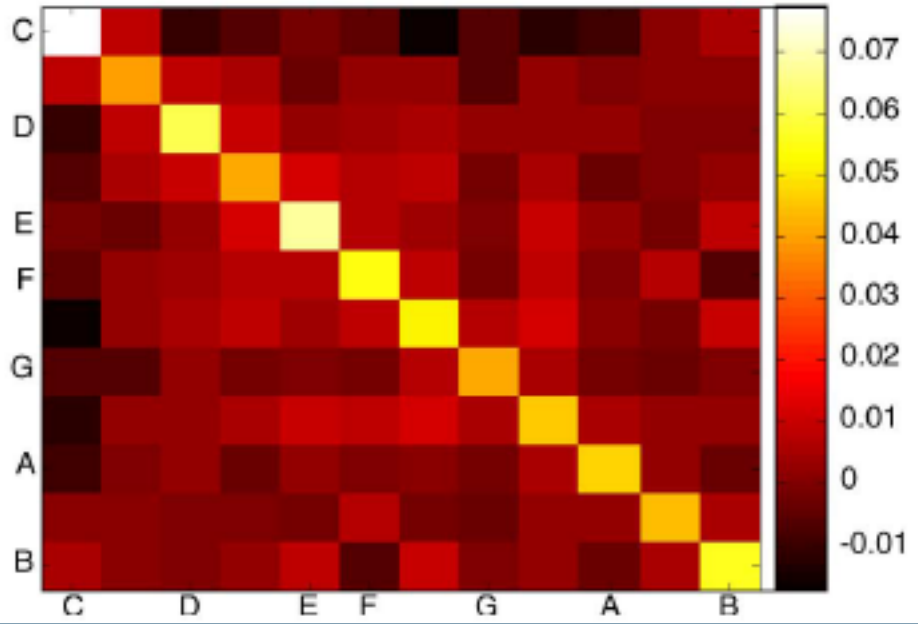
$\mathbf{P}_{\text{tr}}$

HMM training and modelling



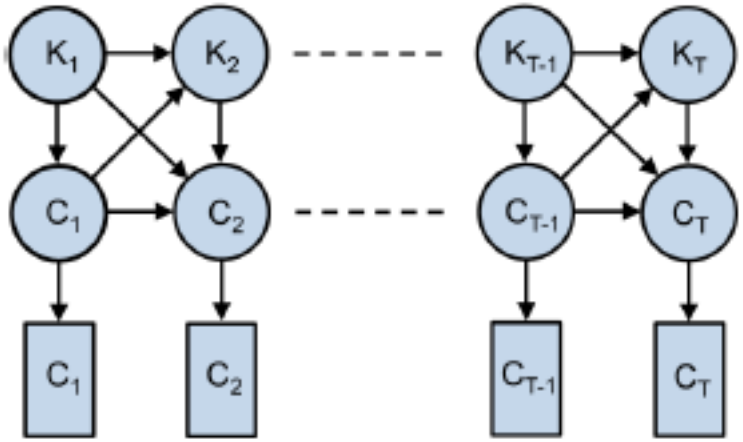
μ

Σ for the C chord



Modeling the key information

Simultaneous estimation of chords and keys can be obtained by including an additional hidden chain into an HMM architecture.



Cons: many more conditional probabilities to be incorporated.

Dynamic Bayesian Networks

Observation: chords depend on higher level musical structure.

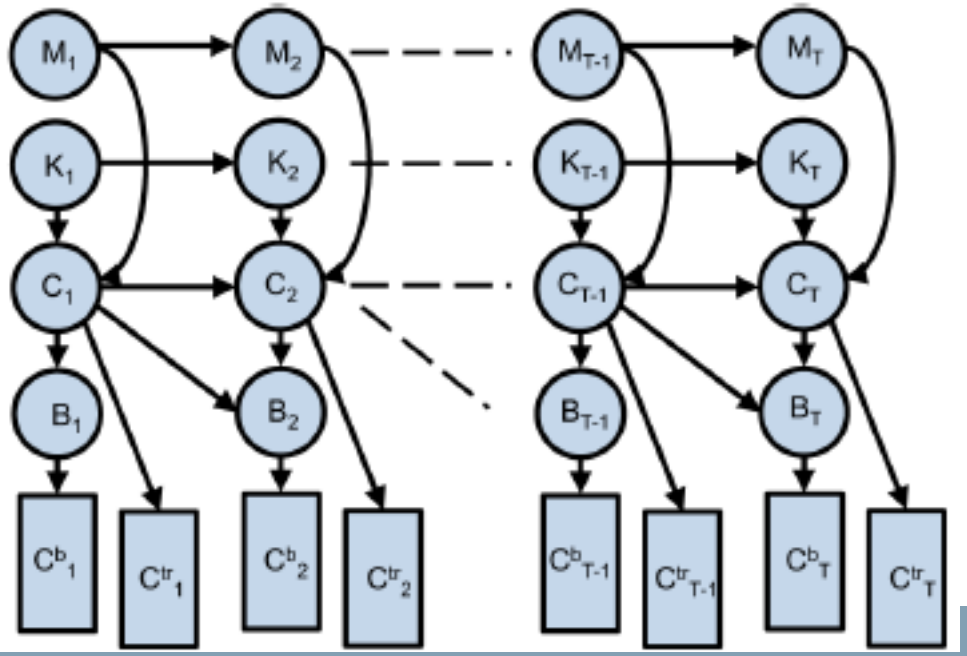
DBN aim at incorporating higher level musical structure.

Hidden variables:

- Metric, musical key, chord and bass note
- Observed variables: treble and bass chromagrams.

Dynamic Bayesian Networks

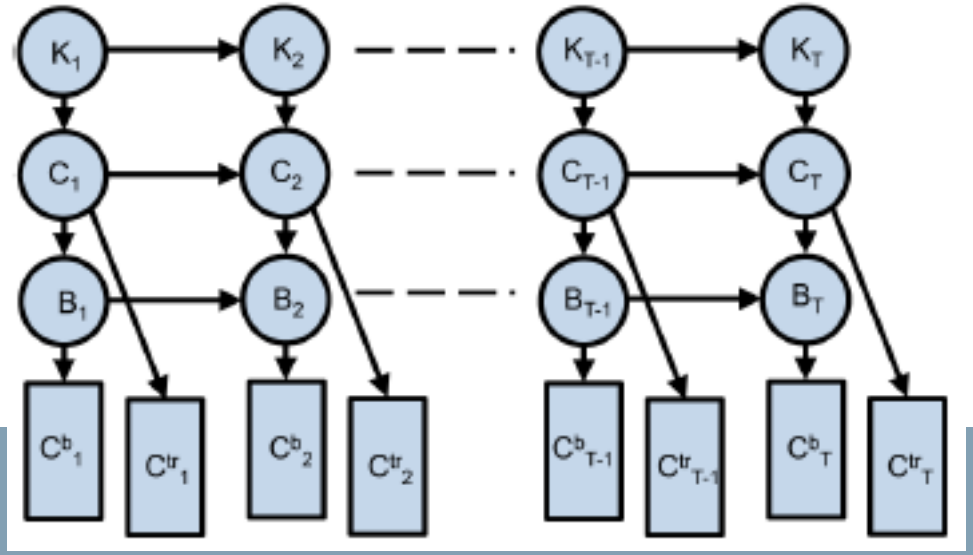
Structure of the Dynamic Bayesian Network [Mauch 2010]:



Harmonic progression Analyzer

Removal of the dependency of the bass note from the chord for chords older than the past one

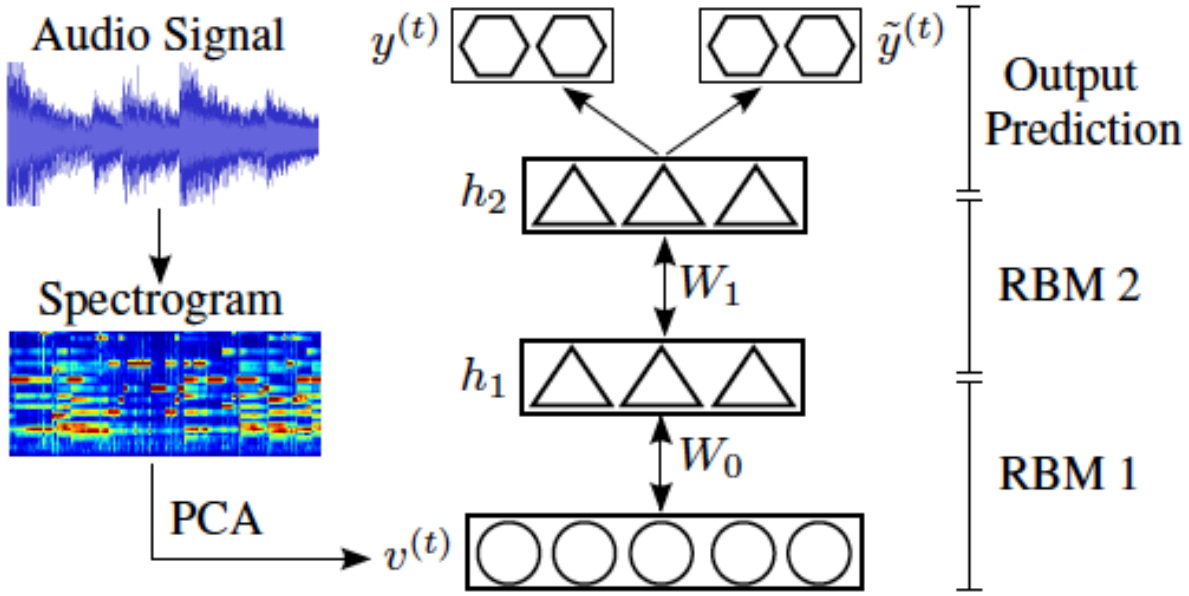
Removal of the metric from the hidden variables set



Recurrent Neural Networks

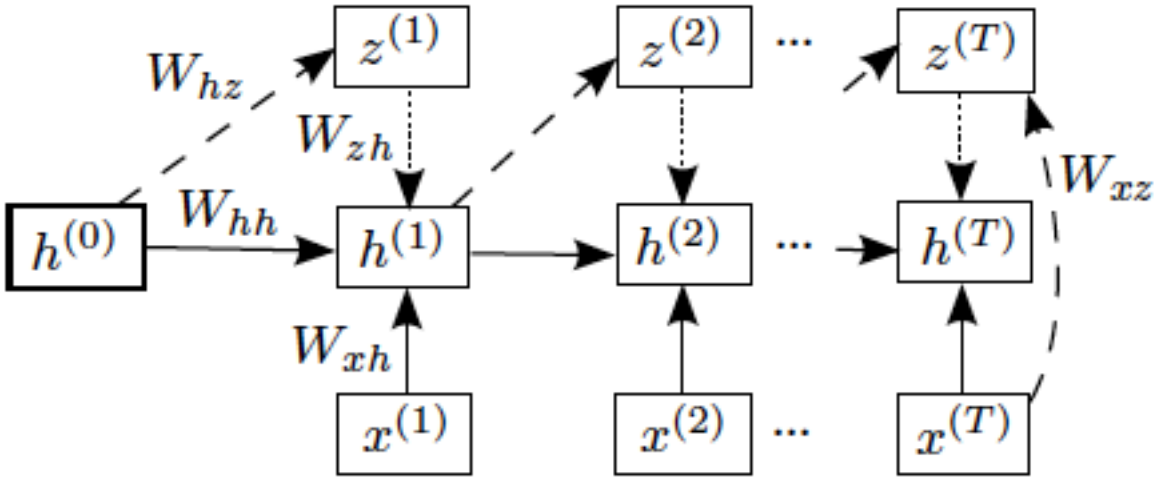
In [Boulanger2013] authors propose an alternative solution, based on Recurrent Neural Networks:

- The input signals are the spectrograms, and a deep belief network is devoted to the inference of features (i.e. hidden layers) relevant for the problem at hand from the spectrogram.



Recurrent Neural Networks

- In a successive stage, once the features have been extracted, chords can be inferred from the features using a RNN.



Note: an interdependence between features at different time instants. This implements the possibility of adding history into the estimation of the current chord.

POLITECNICO
MILANO 1863

REFERENCES

[Harte2010] C. Harte, “Extraction of Harmony Information from Music Signals”, PhD Thesis, Queen Mary University of London, Department of Electronic Engineering

[McVicar2014] M. Mc Vicar, R. Santos-Rodriguez, Y. Ni, T. De Bie, “Automatic Chord Estimation from Audio: a review of the state of the art”

[Brown1991] J. Brown, “Calculation of a Constant–Q spectral transform,” J. Acoust. Soc. Amer., vol. 89, no. 1, pp. 425–434, 1991.

[Ono2008] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram,” in Proc. Euro. Signal Process. Conf., 2008, pp. 445–450.

[Reed2009] J. Reed, Y. Ueda, S. Siniscalchi, Y. Uchiyama, S. Sagayama, and C. Lee, “Minimum classification error training to improve isolated chord recognition,” in Proc. 10th Int. Soc. Music Inf. Retrieval, 2009, pp. 609–614

[Pauws2004] S. Pauws, “Musical key extraction from audio,” in Proc. 5th Int. Soc. Music Inf. Retrieval, 2004, vol. 4, pp. 66–69.

[Lee2006] K. Lee and M. Slaney, “Automatic chord recognition from audio using an HMM with supervised learning,” in Proc. 7th Int. Soc. Music Inf. Retrieval, 2006, pp. 133–137.

[Papadopoulos2007] H. Papadopoulos and G. Peeters, “Large-scale study of chord estimation algorithms based on chroma representation and HMM,” in Proc. Int. Workshop Content-Based Multimedia Indexing, 2007, pp. 53–60.

[Mauch2010] M. Mauch and S. Dixon, “Simultaneous estimation of chords and musical context from audio,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 6, pp. 1280–1289, Aug. 2010.

[Varewyck2008] M. Varewyck, J. Pauwels, and J. Martens, “A novel chroma representation of polyphonic music based on multiple pitch tracking techniques,” in Proc. 16th Int. Conf. Multimedia, 2008, pp. 667–670.

[Ni2012] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie, “An end-to-end machine learning system for harmonic analysis of music,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 6, pp. 1771–1783, Aug. 2012

[Fujishima 1999] T. Fujishima, “Realtime chord recognition of musical sound: A system using Common Lisp Music,” in Proc. Int. Comput. Music Conf., 1999, pp. 464–467

[Bello2005] J. Bello and J. Pickens, “A robust mid-level representation for harmonic content in music signals,” in Proc. 6th Int. Soc. Music Inf. Retrieval, 2005, pp. 304–311

[Bello2011] T. Cho and J. Bello, “A feature smoothing method for chord recognition using recurrence plots,” in Proc. 12th Int. Soc. Music Inf.Retrieval Conf., 2011, pp. 651–656.

[Baum1970]L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164–171, 1970.

[Welch2003] Hidden Markov Models and the Baum–Welch Algorithm, IEEE Information Theory Society Newsletter, Dec. 2003, available at http://www-rcf.usc.edu/~lototsky/MATH508/Baum-Welch.pdf.

[Viterbi1967]A.J. Viterbi "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260-269, April 1967

Boulanger-Lewandowski, Nicolas, Yoshua Bengio, and Pascal Vincent. "Audio Chord Recognition with Recurrent Neural Networks." ISMIR. 2013.