Introduction

* Commercial applications and interests in DSP for music signal processing

What is AMT ?

The capability of transcribing music audio into music notation is a fascinating example of human intelligence. It involves perception (analyzing complex auditory scenes), cognition (recognizing musical objects), knowledge representation (forming musical structures), and inference (testing alternative hypotheses). Automatic music transcription (AMT), i.e., the design of computational algorithms to convert acoustic music signals into some form of music notation, is a challenging task in signal processing and artificial intelligence. It comprises several subtasks, including multipitch estimation (MPE), onset and offset detection, instrument recognition, beat and rhythm tracking, interpretation of expressive timing and dynamics and score typesetting.

*[E. Benetos, S. Dixon, Z. Duan, S. Ewert, “Automatic Music Transcription”, in IEEE SPS Journal Vol. 36 January]*

Applications

A successful AMT system would enable a broad range of interactions between people and music, including music education (e.g., through systems for automatic instrument tutoring), music creation (e.g., dictating improvised musical ideas and automatic music accompaniment), music production (e.g., music content visualization and intelligent content-based editing), music search (e.g., indexing and recommendation of music by melody, bass, rhythm, or chord progression), and musicology (e.g., analyzing jazz improvisations and other nonnotated music). As such, AMT is an enabling technology with clear potential for both economic and societal impact.

*[E. Benetos, S. Dixon, Z. Duan, S. Ewert, “Automatic Music Transcription”, in IEEE SPS Journal Vol. 36 January]*

* Introduce DSP concepts
  + Sampling theorem

The sampling theorem is a consequence of digitizing analogue signals. Sampling an analogue signal stores quantized values of the amplitude of a continuous signal at regular intervals determined by the sampling rate of the ADC system. The result of this conversion will be an array stored in memory with N total samples that represent the analogue signal.

The sampling theorem says that to avoid higher frequency components aliasing as lower frequencies components after sampling the following must be satisfied.

Fs > 2B

Where B is the highest frequency expected in the signal. Frequently a sampling rate of 44.1 kHz is used in audio recording because the range of human hearing is from 20-20kHz.

* + DFT
    - Frequency resolution
    - Zero padding
  + STFT

As in other audio-related applications, the most popular tool for describing the time-varying energy across different frequency bands is the short-time Fourier Transform (STFT), which, when visualized as its magnitude, is known as the spectrogram.

Formally, let x be a discrete-time signal obtained by uniform sampling a waveform at a sampling rate Fs Hz. Using a N-point tapered window w (eg. Hamming w[n] = 0.5-0.46cos(2\*pi\*n/N) for n element of [0: N-1]) and an overlap of half a window length we obtain the STFT

With t element of [0 : T-1] and k element of [0:k]. Here, T determines the number of frames , K = N/2 is the index of the last unique frequency value as dictated by the Sampling Theorem. Thus X(t,k) corresponds to the window beginning at time t\*N/(2\*Fs) in seconds and frequency

in Hz.

*[A. Klapuri and M. Davy, Eds., Signal Processing Methods for Music Transcription. New York: Springer-Verlag, 2006.]*

* + - Windowing

To improve the performance of the STFT several different windowing functions can be used to smooth discontinuities between each frame. In audio recordings this is extremely important to ensure the integrity of the signal is conserved.

* + - Frequency and time resolution trade off

The frequency resolution achieved by an STFT is directly influenced by the window length L and the sampling frequency Fs. The time resolution possible is the inverse of the frequency resolution.

Given that humans can discern changes in a signal as small as 10ms, the frequency resolution has imposed limits.

* + - Log frequency spectrograms

Note that the Fourier coefficients of x are linearly spaced on the frequency axis. Using suitable binning strategies, various approaches switch over a logarithmically spaced frequency axis, by using mel-frequency bands or pitch bands. Keeping the linear frequency axis puts greater emphasis o the high-frequency regions of the signal, thus accentuating the aforementioned noise bursts visible as high-frequency content. One simple yet important step often applied in the processing of music sginals, is referred to as logarithmic compression. Such a compression not only accounts for the logarithmic nature that describes how humans perceive sound but also balances out the dynamic range of the signal.

*[A. Klapuri and M. Davy, Eds., Signal Processing Methods for Music Transcription. New York: Springer-Verlag, 2006.]*

* Introduce AMT and pitch detection

Here, an important question is why a trained musician has no problem in analyzing a chord containing two notes one octave apart. Better understanding in which what two overlapping partials interact and how their amplitude can be precisely estimated is certainly of central importance.

*[C. Yeh and A. Roebel, ‘The expected amplitude of overlapping partials of harmonic sounds’ in Proc. Int. Conf. Acoust., Speech, Signal Process (ICASSP’09), Taipei, Taiwan, 2009, pp. 316-319]*

* Beat detection

The musical aspects of tempo, beat, and rhythm play a fundamental role for the understanding of and the interaction with music. It is the beat, the steady pulse that drives music forward and provides the temporal framework of a piece of music. Intuitively, the beat can be described as a sequence of perceived pulses that are regularly spaced in time and correspond to the pulse a human taps along when listening to the music. The term tempo then refers to the rate of the pulse. The term tempo then refers to the rate of the pulse. Musical pulses typically go along with note onsets or percussive events. Locating such events within a given signal constitutes a fundamental task, which is often referred to as onset detection.

The objective of onset detection is to determine the physical starting times of notes or other musical events as they occur in a music recording. The general idea is to cpaure sudden chanes in the music signal, which are typically caused by the onset of novel events. As a result, one obtains a so-called novelty curve, the peaks of which indicate onset candidates.

Much more challenging is the detection of onsets in the case of non-percussive music, where one often must deal with soft onsets or blurred note transitions. This is often the case for genres dominated by string instruments.

*[A. Klapuri and M. Davy, Eds., Signal Processing Methods for Music Transcription. New York: Springer-Verlag, 2006.]*

The peaks of the novelty curve typically indicate the positions of note onsets. Therefore, to explicitly determine the positions of note onsets, one employs peak picking strategies based on fixed or adaptive thresholding.

*[J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, ‘A tutorial on onset detection in music signals’, IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 1035-1047, Sep. 2005]*

*[R. Zhou, M. Mattavelli, and G. Zoia, ‘Music onset detection based on resonator time frequency image’, IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 8, pp. 1685-1695, Nov. 2008]*

* Polyphonic voices

Timbre is defined as the ‘attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar’.

*[USA Standard Acoustical Terminology American National Standards Inst., Tech. Rep. S1.1-1960, 1960]*

Human listeners, especially trained musicians, can switch between a ‘holistic’ listening mode where they consider a music signal as a coherent whole, and a more analytic mode where they focus on the part played by a particular instrument.

*[R. Erickson, Sound Structure in Music. Berkley, CA : Univ. of California, 1975]*

* MPE

Given the extensive literature of speech signal analysis, it seems natural that numerous signal processing studies have focused on monophonic signals. While monophonic signals certainly result in better performances, in recent years there has been an intensifying gradual focus on the more challenging and realistic case of polyphonic music.

The problem is indeed particularly difficult for music signals for which concurrent notes stand in close harmonic relation. For extreme cases such as complex orchestral music, where one has a high level of polyphony, multipitch estimation becomes intractable with todays methods. B

* Source separation

A common and relatively successful approach is to split up the polyphonic signal into individual components that are individually processed as monophonic signals. This process is known as source separation as is a highly related topic to AMT.

*[A. Klapuri and M. Davy, Eds., Signal Processing Methods for Music Transcription. New York: Springer-Verlag, 2006] – pg. 13 section Polyphony and Musical Voices*

Three main situations occur in source separation problems. The determined case corresponds to the situation where there are as many mixture signals as different sources in the mixtures. Contrary, the overdetermined case refers to the situation where there are more mixtures then sources. The underdetermined source separation is the most common and most challenging case. The problem of source separation classically includes two major steps that can be realized jointly : estimating the mixing matrix and estimating the sources. Let X = [x\_1(n), … x\_N(n)]^T be the N mixture signals, S = [ s\_1(n),…., s\_M(n)]^T the M source signals and A = [a\_1, a\_, …, a\_N]^T the NxM mixing matrix with mixing gains a\_i = ( a\_i1, a\_i2….a\_iM). The mixture signals are then obtained by X = AS. This readily corresponds to the instantaneous mixing model.

*[T. Virtanen, ‘Unsupervised learning methods for source separation in monaural music signals’ in Signal Processing Methods for Music Transcription, A. Klapuri and M. Davy, Eds. New York : Springer, 2006, ch.6 pp. 267-296]*

A wide variety of approaches exist to estimate the mixing matrix and rely on techniques such as Independent Component Analysis (ICA), sparse decompositions or clustering approaches. Once the mixing matrix is known the sources can be recovered using heuristic methods, minimization criteria on the error, or time-frequency masking approaches.

* State of the art approaches and challenges faced
  + NMF
  + NN
  + Traditional Signal Processing Methods

System Design

* Libraries used with analysis
* Main approaches
* System block diagram illustrating filters in even library functions such as librosa

Results

* Accuracy, precision and recall results of pitch detection
* STFT graphs showing predictions
* NMF
* NN
* Traditional Signal Processing Methods (CEPSTRUM, ACF) etc.

Future Work

* Areas to improve upon
* Problems unsolved
* Most promising avenues for further research