MELODY EXTRACTION FOR MIREX 2016 USING DYNAMIC MODE DECOMPOSITION

Leonid Pogorelyuk and Clarence Rowley
Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
leonidp@princeton.edu, cwrowley@princeton.edu

ABSTRACT

This abstract describes our system for the audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX) 2017. Our experimental and novel approach involves frequency extraction from audio signals via Dynamic Mode Decomposition (DMD) [1] instead of the commonly used Fourier and Wavelet Transforms. Melody pitch tracking is based on a set of heuristic rules applied to the DMD features extracted from the audio signal.

1. INTRODUCTION

Dynamic Mode Decomposition (DMD) is a data driven method which estimates the eigenvalues of linear autonomous dynamical systems from linear observables of the system [1]. A finite number of simple harmonic oscillators form a linear finite dimensional system and can approximate a musical instrument playing a note. Therefore, DMD can be used to estimate the fundamental frequencies and their harmonics present in a short sample of an audio signal, assuming that it is a superposition of simple harmonic oscillators.

Performing DMD on short time frames of the audio signal is conceptually similar to a short-time Fourier transform (STFT). The resulting DMD features are frequency/magnitude pairs and can be directly used for pitch contour detection [3], forgoing the step of peak detection. Below, we briefly present the DMD based algorithm for frequency detection and the heuristics used in our system for filtering out non-fundamental frequencies.

2. DYNAMIC MODE DECOMPOSITION BASED FREQUENCY DETECTION

An audio signal can be represented by a series of measurements $\{y_t\}$ at constant time intervals $t = k\Delta t$. We further assume that over a short time frame, the signal is formed by superimposing N simple oscillators with angular frequencies $\{\omega_j\}_{j=1}^N$ and complex amplitudes $\{A_j\}_{j=1}^N$:

$$y_t = \mathrm{Re}\left(\sum_{j=1}^{N} A_j \exp\left(i\omega_j k\Delta t\right)\right) \tag{1}$$

It can be shown (paper pending publication) that in order to detect N frequencies, delayed observables with at least $2N$ time delays must be formed:

$$z_t = \begin{bmatrix} y_t & y_{t+1} & \cdots & y_{t+2N-1} \end{bmatrix}^T \tag{2}$$

and the size of the frame must be at least $3N$. This gives a set of measurements which can be arranged as columns of two matrices:

$$\begin{aligned} Z_t &= \begin{bmatrix} z_t & \cdots & z_{t+N-1} \end{bmatrix} \\ Z_{t+1} &= \begin{bmatrix} z_{t+1} & \cdots & z_{t+N} \end{bmatrix} \end{aligned} \tag{3}$$

.

A standard DMD procedure is then to compute the matrix

$$K_t = Z_{t+1} Z_t^+ \tag{4}$$

where Z_t^+ denotes the Moore-Penrose pseudoinverse of Z_t computed via truncated Singular Value Decomposition (SVD).

The eigenvalues $\{\lambda_j\}$ of the square matrix K_t are zeros and $\{\exp\left(\pm i\omega_j \Delta t\right)\}$, within numerical errors. Moreover, the amplitudes $\{A_j\}_{j=1}^N$ (and their conjugates) are approximately the coefficient of the data vector z_t expressed in the basis formed by the eigenvectors of K_t.

The feature extraction step of our system consists of performing DMD with delayed observables over short time frames, and the resulting features at each such frame are pairs of frequency and magnitude:

$$\{(f_j, M_j)\} = \left\{\left(\frac{1}{2\pi\Delta t}\mathrm{Im}\left(\log\lambda_j\right), |A_j|\right)\right\} \tag{5}$$

Figure 1 shows a one-second excerpt from Bach's 8th Invention in F major [2]. DMD features (circles) were computed over 100 millisecond windows at 10 millisecond intervals. High magnitude features correspond to fundamental frequencies and harmonics of the notes being played.

3. PITCH CONTOUR DETECTION AND MELODY SELECTION

Given DMD features we proceed with a contour detection step similar to [3]. The features are processed in order of

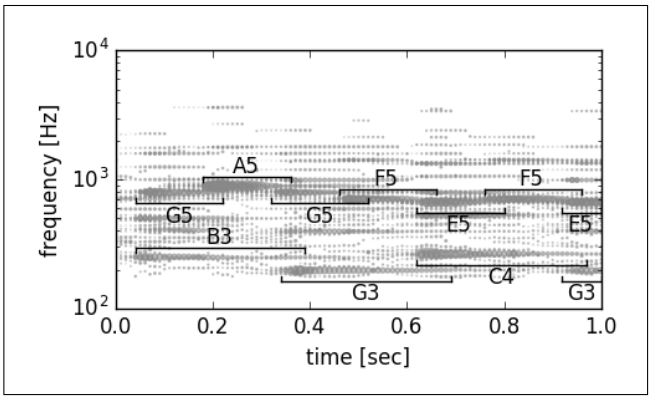

Figure 1. DMD frequency/magnitude pairs (circles) extracted from 100 millisecond frames (the radius of the circle is proportional to the magnitude). The fundamental frequencies of the notes being played are marked by annotated black lines.

decreasing magnitude, until all of them belong to a certain contour. The first feature pair in a new contour defines the fundamental frequency and features at adjacent frames are added if they have close frequencies or are close to the F harmonic of the fundamental frequency.

A contour is then given by a set $\{(f_k, M_k, t_k, F_k)\}$ of feature pairs together with their timing t and harmonic F. The salience C_s and pitch of the contour C_p at time t can be computed via:

$$\begin{aligned} C_s(t) &= \sum_{t_k=t} W_{F_k} M_k \\ C_p(t) &= \frac{1}{C_{s(t)}} \sum_{t_k=t} W_{F_k} M_k \frac{f_k}{F_k+1} \end{aligned} \tag{6}$$

where W_{F_k} are harmonic-based weights that depend on the note represented by the contour. The onset and offset of a contour can be detected via a voicing-detection rule based on the salience [3].

Some of the detected contours correspond to a fundamental frequency of a note, however most of them are either higher harmonics or noise. At the last stage contours are filtered based on their salience and presence of other contours. This process can be summarized as:

1. Sort all contours by increasing pitch
2. If the salience of the contour is above a certain threshold:
   (a) Add the contour to the list of detected notes
   (b) Penalize salience of all contours that have pitches that are harmonics of the current contour's pitch
3. Move to the next lowest pitch contour
4. Repeat steps 2–3

The heuristic penalties incurred at step 2b above depend on the harmonic, the fundamental frequency and the temporal overlap between the two contours.

4. REFERENCES

[1] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz: "On dynamic mode decomposition: theory and applications," *arXiv preprint*, arXiv:1312.0041, 2013.

[2] J. S. Bach and G. Gould "Invention No. 8 in F Major, BWV 779." *The Two and Three Part Inventions.* Sony Classical Records, CD, 1992.

[3] J. Salamon and E. Gómez. "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio. Speech. Lang. Processing*, 20(6):17591770, 2012.