Multiple Fundamental Frequency Estimation of Piano Signals via Sparse Representation of Fourier Coefficients

Cheng-Te Lee **Yi-Hsuan Yang** **Keng-Sheng Lin** **Homer Chen**

National Taiwan University

aderleee@gmail.com affige@gmail.com pridek0912@gmail.com homer@cc.ee.ntu.edu.tw

ABSTRACT

In this paper, multiple fundamental frequency estimation of piano signals is formulated as a sparse representation problem, for which the lowest and the highest possible values of fundamental frequencies are estimated first. Then, under the assumption that the waveforms of the piano notes are pre-stored and that the Fourier coefficients of a given music signal can be represented as a linear combination of the Fourier coefficients of the prestored waveforms of piano notes, we solve the sparse representation problem by L1 minimization, followed by a temporal smoothing based on the hidden Markov models.

1. PROPOSED METHOD

Multiple fundamental frequency estimation entails the determination of onset time, offset time, and fundamental frequency of each note of a music signal. It is crucial for content-based music information retrieval. In this work, we limit the input signal to that produced by piano. The proposed method consists of three steps: estimation of the fundamental frequency bounds, calculation of the sparse representation coefficients, and temporal smoothing. The first two steps are performed in a frame-based manner with 10-millisecond hop size between successive frames [1], [3] and each frame being 100 milliseconds long. Then in the final step we apply temporal smoothing to the resulting fundamental frequencies. The detail of each step is described below.

1.1 Estimation of Bounds of Fundamental Frequency

The first step of our proposed method estimates the lowest and highest values of the fundamental frequencies of an input frame to reduce the search space of the sparse representation coefficients. In this way, we can reduce the time complexity of the subsequent operations for sparse representation calculation.

First, we obtain frequency-domain information (more specifically, Fourier coefficients) by applying the short-time Fourier transform (STFT) to the input frame. Then the local maxima of the Fourier coefficients are derived by using the method proposed in [5]. Under the assumption that the Fourier coefficient of a fundamental frequency is a local maximum in the frequency domain and larger than the Fourier coefficients of its harmonics, irrelevant local maxima are eliminated as follows:

- Consider every frequency whose Fourier coefficient is a local maximum as a candidate fundamental frequency.
- Rule out a candidate frequency if it is an integer multiple of another candidate but its coefficient is smaller.

The lowest and highest frequencies of the remaining candidates set the search range of the fundamental frequency of the corresponding frame.

1.2 Sparse Representation

We assume that the Fourier coefficients of an input frame are a linear combination of the Fourier coefficients of pre-stored waveforms of individual piano notes [1]. Denote the dictionary for the ith key of piano by $\mathbf{A}_i = [\mathbf{a}_{i,1}|\mathbf{a}_{i,2}|\ldots|\mathbf{a}_{i,n_i}]$, where $\mathbf{a}_{i,k}$ is a column vector containing the Fourier coefficients of the kth short-time segment of the pre-stored waveform of that key and n_i is the number of segments. We define the matrix $\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2|\ldots|\mathbf{A}_{88}]$, which contains the Fourier coefficients of segments of pre-stored waveforms [2], [3]. Let the column vector $\mathbf{y}$ be the Fourier coefficients of an input frame, the problem now is to find a sparse coefficient vector $\mathbf{x}^*$ such that

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \| \mathbf{x} \|_0 \text{ subject to } \mathbf{y} = \mathbf{A}_\mathbf{p}\mathbf{x} \qquad (1)$$

where $\mathbf{A}_\mathbf{p} = [\mathbf{A}_l|\mathbf{A}_{l+1}|\ldots|\mathbf{A}_h]$ and the bounds of the fundamental frequency derived from the first step correspond to the lth and the hth keys of piano. Because finding the solution of (1) is NP-hard, we reformulate it as the following constrained L1 minimization problem,

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \| \mathbf{x} \|_1 \text{ subject to } \mathbf{y} = \mathbf{A}_\mathbf{p}\mathbf{x} \qquad (2)$$

After obtaining the sparse coefficient vector $\mathbf{x}^*$ of (2), we consider that a note (a fundamental frequency) is present in an input frame if the summation of the coefficients corresponding to that note is larger than a predefined threshold.

1.3 Temporal Smoothing by Hidden Markov Models

The STFT approach described above treats the short-time frames independently, leaving the temporal structure of music unexploited. To address this issue, we use two-state (on and off) hidden Markov models (HMMs) to model each note independently [4]. For each note, we want to maximize

$$\prod_t p(x_t \mid q_t) p(q_t \mid q_{t-1}) \qquad (3)$$

where q_t is the state at time t, x_t is the frame beginning at time t, $p(x_t|q_t)$ is the probability of x_t being observed given q_t, and $p(q_t|q_{t-1})$ is the transition probability between states. Although we do not know $p(x_t|q_t)$, from the conditional probability, we have

$$p(q_t \mid x_t) \propto p(x_t \mid q_t) p(q_t) \tag{4}$$

Therefore, we can maximize

$$\prod_t \frac{p(q_t \mid x_t)}{p(q_t)} p(q_t \mid q_{t-1}) \tag{5}$$

instead of (3). The sparse representation coefficient can be seen as an approximation of $p(q_t|x_t)$. Both the prior $p(q_t)$ and the state transition probability $p(q_t|q_{t-1})$ can be learnt from the training data. We can apply the Viterbi algorithm to find the solution of (5).

After temporal smoothing by HMMs, we obtain the final result.

2. ACKNOWLEDGMENTS

This work was supported by the National Science Council of Taiwan under the contract number NSC 97-2221-E-002-111-MY3.

3. REFRENCES

[1] J. P. Bello, L. Daudet, and M. Sandler, "Time-domain polyphonic transcription using self-generating databases," *in Proc. Convention of the Audio Engineering Society*, 2002.

[2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[3] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification via sparse representations of auditory temporal modulations," *in Proc. European Signal Processing Conference*, 2009.

[4] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Advances in Signal Processing*, vol. 8, pp. 1–9, 2007.

[5] O. Lartillot, P. Toiviainen and T. Eerola, MIR toolbox, https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox