ICSV 21
13-17 July, 2014, Beijing/China
The 21st International Congress on Sound and Vibration
13-17 July, 2014, Beijing/China

MULTI-PITCH ONSET DETECTION VIA TEMPORAL SEGMENTATION AND SEGMENTAL ANALYSIS

Yulong Wan, Xianliang Wang, Ruohua Zhou, Yonghong Yan

Institute of Acoustics, Chinese Academy of Sciences, Beijing, China 100190,
e-mail: {wanyulong,wangxianliang,zhouruohua,yanyonghong}@hccl.ioa.ac.cn

This paper presents a novel multi-pitch onset detection approach, which consists of a temporal segmentation stage and segment-based onset detection stage. Unlike most conventional approaches, the proposed method first exploits obvious onsets by adaptively matched filtering on the amplitude envelope, and segments the original audio stream into small clips using the detected onsets. Then the onsets in each clip are obtained on pitch energy spectrum, which is the weighted harmonic summation of average energy spectrum calculated from resonator time-frequency image. As the matched filtering and following processes are directly related with input data, no prior knowledge or statistical information is thereby required. Experiments show a significant improvement over existing state-of-the-art methods.

1. Introduction

Polyphonic music transcription is a process of converting an audio waveform into a parametric representation, and has been long seemed as an extremely hard task. This parametric representation includes the musical notes, their pitches, starting times, and durations. The task of recovering the start times of musical notes is known as note onset detection. A successful onset detection method enables the temporal segmentation of an audio signal at a meaningful time-scale. Within music information retrieval (MIR) research, onset detection forms the basis of many higher level processing tasks, including beat tracking and interactive musical accompaniment.

During the past decade, many solutions for onset detection have been proposed, most of which consist of three stages illustrated in Fig. 1 (B). First, the signal is preprocessed using noise reduction, time-frequency analysis, etc. A mid-level representation, referred to as an onset detection function, is then extracted from the pre-processed result. The aim of the onset detection function is to exhibit peaks at likely onset locations by measuring changes in the short term properties of the audio signal; for example: energy, high frequency content [1], phrase information [2], or pitch [3]. For a review of feature types, readers can refer to [4] and [5]. Once the onset detection function has been generated, the temporal locations of the note onsets can be recovered by applying a peak-picking algorithm.

However, these methods mentioned above usually apply analysis on the whole file duration, which incurs a high computational cost and system pressure in a short time. Also, the whole-file level processing may lose some details which leads to high miss-rate of onsets.

Keeping the above problems in mind, we propose a novel onset detection combining automatic temporal segmentation and segmental onset detection. This is inspired by the observation that the onsets of amplitude envelope always coincide with some onsets of music notes. In Fig. 1, (A) and (C) show the flows of temporal segmentation and segmental onset detection, respectively.

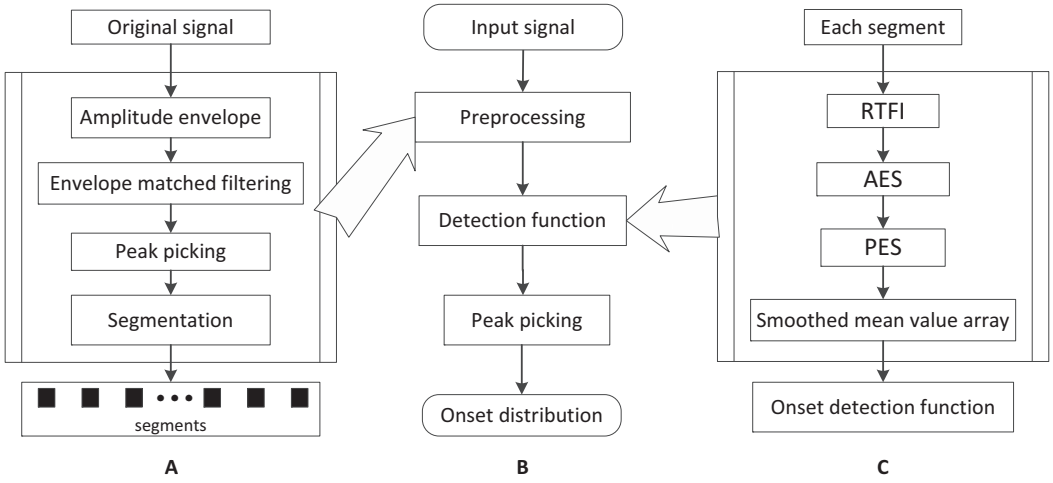

Figure 1. System flow of the proposed onset detection method.

The remainder of this paper is as follows. Section 2 describes the envelope onset detection approach based on adaptive matched filtering in detail. In Sec. 3, we present the onset detection method in each segment. Section 4 describes the experiments for performance evaluation. Section 5 concludes this paper.

2. Envelope onset detection

In signal processing, the matched filtering is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal [6]. This is equivalent to convolving the unknown signal with a conjugated time-reversed version of the template. For a discrete-time system, the matched filtering can be expressed as follows in Eq. (1).

$$y[n] = \sum_{k=-\infty}^{\infty} h[n-k]x[k]. \tag{1}$$

where $x[n]$ is the observed signal, $h[n]$ is a linear matched filter, and $y[n]$ is the output response. To detect the onsets of notes, we need to seek a filter h to maximize the output around the area of onsets.

Conventional onset detection algorithms usually use the variation of the envelope amplitude or its energy to judge whether there is an onset, just as the Eq. (2), which is equivalent to the convolution of A_k with the following filter F_d in Eq. (3).

$$D_k = A_k - A_{k-1}. \tag{2}$$

$$F_d = [1, -1]. \tag{3}$$

However, in practice, observing the variation of the envelope amplitude directly will often lead to misjudgment. In [7], Pauws proposed an onset detection named “surf method”, which used the slope of the envelope to detect the onsets. The slope is determined using five adjacent time slots, which is in fact equivalent to the convolution of A_k with the following filter F_s in Eq. (4).

$$F_s = [2, 1, 0, -1, -2]/10. \tag{4}$$

In [8] and [9], matched filters are applied instead of Eq. (3) and Eq. (4) for onset detection. The results show that the matched filter method is very suitable for the onset detection of fixed-type audio signals, such as human voice, music of piano and other specific musical instruments, etc. However,

these two methods both use fixed matched filters, which are obtained from statistics and can not be changed adaptively according to the analyzed file itself.

In this paper, we consider obtaining a matched filter by summarizing the envelope amplitudes around obvious peaks. As the sample number of each frame is the same with that in [8], the length of the matched filter is also fixed to 12.

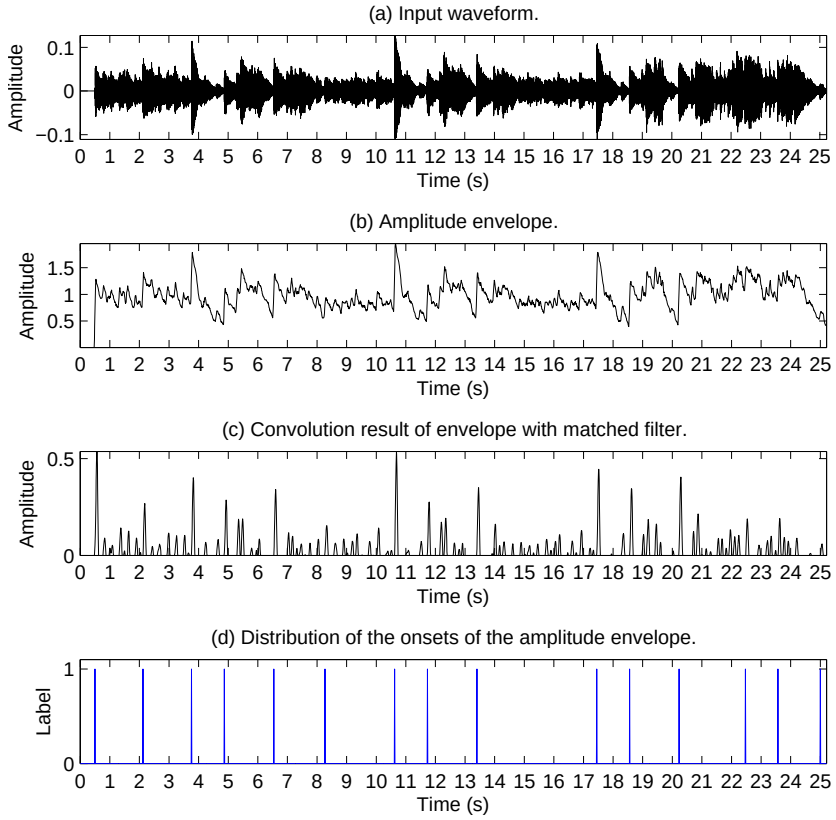

Figure 2. Onset detection of amplitude envelope.

The proposed onset detection algorithm for amplitude envelope is detailed as follows:

1) Find the envelope amplitude A_o for each time frame as Eq. (5).

$$A_o[k] = \max_{kn_0<n<(k+1)n_0}(|x[n]|). \tag{5}$$

where x represents the input signal, n_0 is the sample number in each frame, and k is the frame index.

2) Perform the normalization as Eq. (6) to obtain a more objective onset detection result.

$$A_n[k] = \frac{A_o[k]}{mean(A_o)}. \tag{6}$$

where $mean(A_o)$ denotes the mean value of all the amplitudes.

3) To avoid the miss-detection of the onsets with lower amplitudes, we take the fractional power of the normalized amplitude envelope A_n by Eq. (7) as follows.

$$A_f[k] = (A_n[k])^{\kappa}, \quad 0 < \kappa < 1. \tag{7}$$

4) Smooth the fractional power amplitude by simply median filtering just as Eq. (8).

$$A_s[k] = \frac{1}{2 \times n_s + 1} \sum_{n=k-n_s}^{n=k+n_s} A_f[n]. \tag{8}$$

where n_s denotes the number of time frames around used for smoothing.

5) Pick the obvious peaks in A_s, and store the labels in P_e as Eq. (9), in which 1 for "peak", and 0 for "non-peak". The $P_e[k]$ is labeled as 1 only if A_s satisfies the following rules:

$$\textbf{if } A_s[k] > A_s[k_r],\ \forall k_r \in (k-7, k+4),\ k_r \neq k,\ \textbf{then } P_e[k] = 1;\ \textbf{else } P_e[k] = 0. \tag{9}$$

6) Obtain the matched filter by summarizing all the envelope around the peaks as Eq. (10).

$$MF[k] = \frac{1}{N_p} \sum_{n=1}^{N_p} A_s[k_n - 8 + k], \quad 1 \leq k \leq 12. \tag{10}$$

where k_n is the frame index of the n-th peak, k denotes the frame index of the matched filter and N_p is the number of peaks in all.

7) Perform the convolution of A_s and the matched filter as the following Eq. (11).

$$C[k] = \max(\sum_{\tau=1}^{12} A_s[k-\tau] \cdot MF[\tau], 0). \tag{11}$$

8) Use simple peak picking to get the envelope onsets as the Eq. (12).

$$\textbf{if } (C[k] > C[k_r]) \cap (C[k] > C_{thre}),\ \forall k_r \in (k-3, k+3),\ k_r \neq k, \quad \textbf{then } (k-6)n_0 \text{ is a onset location.} \tag{12}$$

After obtaining the onsets of amplitude envelopes, we can divide the input signal into small segments with each onset as a beginning point of a new segment, then detect the onsets in each segment, which can be combined in the end to yield the final results. Figure 2 shows the processing sequence of the envelope onset detection.

3. Segmental onset detection

In this section, the details of onset detection in each segment are presented. We employ a computationally efficient time-frequency representation named Resonator Time-Frequency Image (RTFI), which was proposed by Zhou in [10]. The RTFI has been proved very fit for music signal analysis in [3], [8], and [11]. It selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis just as follows in Eq. (13). Using the RTFI, different time-frequency resolutions can be selected by simply setting a few parameters.

$$RTFI(t,\omega) = s(t) * I_R(t,\omega) = r(\omega) \int_0^t s(\tau) e^{r(\omega)(\tau-t)} e^{-j\omega(\tau-t)} d\tau \tag{13}$$

where $I_R(t,\omega) = r(\omega)e^{(-r(\omega)+j\omega)t}$, $t > 0$, denotes the impulse response of the first-order complex resonator filter with oscillation frequency ω. The factor $r(\omega)$ before the integral in Eq. (13) is used to normalized the gain of the frequency response when the resonator filter's input frequency is the oscillation frequency. The details of RTFI can be referred to [10]. In this paper, we employed the constant-Q RTFI, as the inter-harmonic spacings are the same for any periodic sounds. The frequency band of each semitone is covered using one filter. The number of frequency bins per octave is 12, for the reason that each octave has 12 semitones. A total of 89 filters are used to cover the analyzed frequency range for the entire 88 music notes of piano, which extends from 25.96 Hz to 4.43 kHz.

To reduce the memory usage of storing the RTFI values, the RTFI is separated into different time frames, and the Average RTFI Energy Spectrum (AES) is calculated in each frame as in Eq. (14):

$$AES(k,\omega_m) = db(\frac{1}{M}\sum_{n=(k-1)\times M+1}^{k\times M}|RTFI(n,\omega_m)|^2) \tag{14}$$

Where k denotes the frame index, $db(\cdot)$ converts the value to decibels, M is an integer equal to the number of samples in each frame. $RTFI(n,\omega_m)$ represents the value of the discrete RTFI at sampling point n and frequency ω_m. The detailed description of the discrete RTFI can also be found in [10]. The AES is then used as the input for our segmental onset detection.

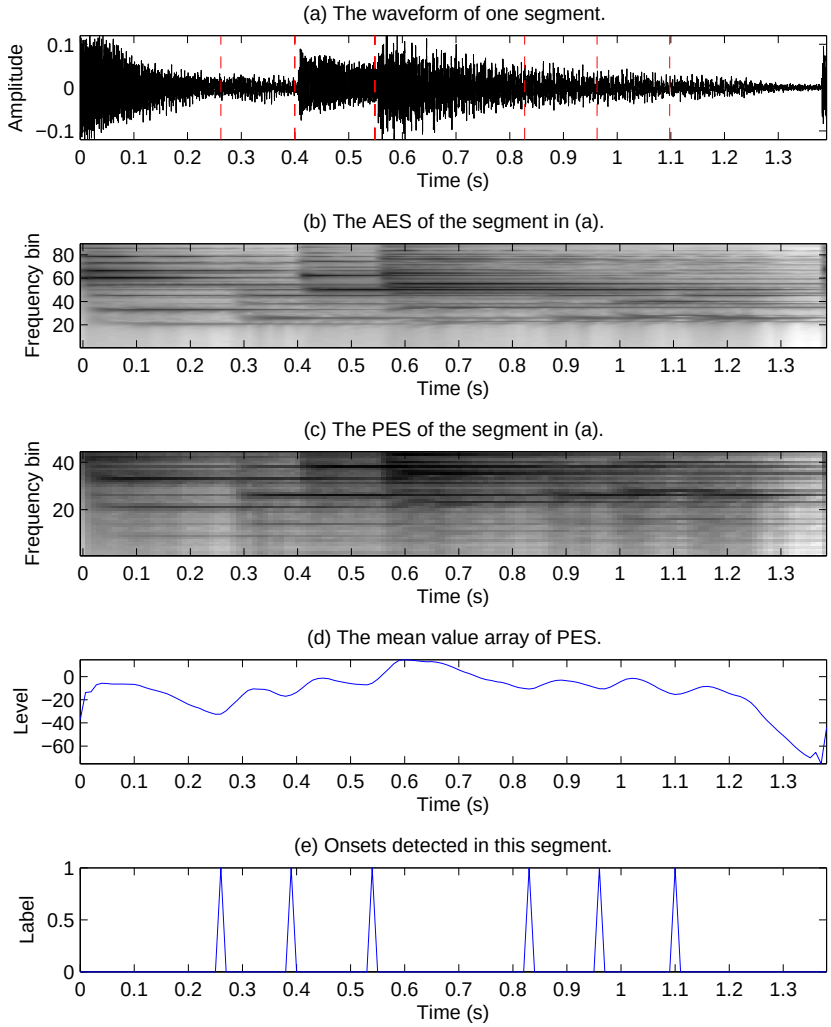

Figure 3. Onset detection in each segment.

After obtaining the AES of each segment, we first calculate the pitch energy spectrum (PES) using weighted harmonic summation in Eq. (15), which is similar to the pitch estimation method proposed in [12]. The AES and PES are illustrated in Fig. 3 (b) and (c), respectively.

$$PES(k,\omega_m) = \sum_{n=1}^{N_h} h^{n-1} AES(k, n\omega_m) \tag{15}$$

where h is a compression factor, $0 < h < 1$, N_h is the maximum number of harmonics, and m is the frequency bin of PES, $1 < m < N_m/2$. Here, $N_m = 89$.

Then we calculate the mean value array of PES as follows in Eq. (16), which can be seen in Fig. 3 (d), and use the same peak picking method as Eq. (12) to detect the onsets in each segment. The onset detection result of the segment example can be seen in Fig. 3 (e). The segment sample is illustrated in Fig. 3 (a), in which the red dashed lines denotes the reference onsets of notes in this segment. We can see that this segmental analysis can improve the onset detection by finding out the ones that can't be detected using temporal methods.

$$MA(k) = \frac{2}{N_m} \sum_{n=1}^{N_m/2} PES(k, n\omega_m) \tag{16}$$

The last phrase is to combine all the onsets detected in each segment to yield the overall result. First the onsets from all segments are sorted in time order and regarded as onset candidates hereafter. Each candidate is assigned a loudness value calculated from Eq. 16. Then we drop out candidates that are too close (< 100 ms) to a louder one. Among equally loud but too close ones, the middle one is chosen and the others are abandoned. The remaining onset candidates are accepted as true ones.

4. Performance evaluation

4.1 Experiment database and evaluation metrics

We carried out experiments on the a small database chosen randomly from the MUS subset of MAPS database. The test set contains 9 pieces sampled at 44.1 kHz, 16-bit, which consists of stereo ones under 9 different recording conditions named as "AkPnBcht", "AkPnBsdf", "AkPnCGdD", "AkPnStgb", "ENSTDkAm", "ENSTDkCl", "SptkBGAm", "SptkBGCl", and "StbgTGd2". Details of these conditions can be found in [13]. The onset locations have been adjusted by the creators. There are totally 18146 onsets in our test set, and the onset number in each file can be found in Table 2.

To evaluate the performance of onset detection, the detected onsets must be compared with the reference ones. For a onset detected at time t, if there is a reference one within a tolerance time-window $[t - 50ms, t + 50ms]$, it is considered to be true positive (TP), otherwise, it is false positive (FP). The reference onsets outside all the tolerance windows are counted as false negative (FN). Here, we use the Precision, Recall and F-measure to summarize the results, as expressed in Eq. (17):

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}}. \qquad Recall = \frac{N_{TP}}{N_{TP} + N_{FN}}.$$
$$\textit{F-measure} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{17}$$

Where N_{TP} is the number of TP onsets (or TP notes) detected, N_{FP} is the number of FP ones, and N_{FN} is the number of FN ones.

4.2 Results

In this paper, the used parameters determined by preliminary experiments are listed in Table 1. And the results are compared with the baseline system proposed by Zhou in [3].

Table 1. Parameters used in the implementation of the proposed onset detection method.

$n_0 = 441$	$\kappa = 0.5$	$n_s = 2$	$M = 441$	$h = 0.7$	$N_h = 5$

First, we evaluate the performance of envelope onset detection proposed in Sec. 2. Since the purpose of this stage is to detect obvious onsets for temporal segmentation, we are more concerned about precision. From the results in Table 2, we could see that the proposed envelope onset detection can detected some note onsets with high accuracy which could be used for temporal segmentation.

Table 2. Performance of the proposed envelope onset detection in 9 different recording conditions.

Filename_Condition	Number	Precision	Recall	F-measure
MUS-alb_se3_AkPnBcht	1291	1.000	0.113	0.203
MUS-alb_se3_AkPnBsdf	1291	0.995	0.249	0.398
MUS-bk_xmas5_AkPnCGdD	1680	0.995	0.259	0.411
MUS-appass_3_AkPnStgb	3980	0.997	0.159	0.274
MUS-chpn_op35_1_ENSTDkAm	2054	0.982	0.086	0.158
MUS-grieg_butterfly_ENSTDkCl	716	0.958	0.135	0.237
MUS-chpn-e01_SptkBGAm	1243	0.991	0.094	0.172
MUS-appass_1_SptkBGCl	4011	0.983	0.157	0.270
MUS-mz_311_1_StbgTGd2	1880	1.000	0.149	0.259

Then, we give the comparison results of the whole onset detection methods in Table 3, from which we could find that our proposed onset detection method can get much higher Recall than the baseline one, leading to better F-measure in all 9 recording conditions.

Table 3. Performance comparisons of overall onset detection in 9 different recording conditions.

Filename_Condition	System	Precision	Recall	F-measure
MUS-alb_se3_AkPnBcht	Baseline	0.999	0.736	0.847
	Proposed	0.975	0.891	0.931
MUS-alb_se3_AkPnBsdf	Baseline	0.996	0.736	0.846
	Proposed	0.916	0.849	0.881
MUS-bk_xmas5_AkPnCGdD	Baseline	1.000	0.673	0.805
	Proposed	0.953	0.781	0.859
MUS-appass_3_AkPnStgb	Baseline	0.997	0.678	0.808
	Proposed	0.966	0.759	0.850
MUS-chpn_op35_1_ENSTDkAm	Baseline	0.949	0.754	0.840
	Proposed	0.863	0.856	0.859
MUS-grieg_butterfly_ENSTDkCl	Baseline	0.937	0.711	0.809
	Proposed	0.939	0.770	0.846
MUS-chpn-e01_SptkBGAm	Baseline	1.000	0.472	0.642
	Proposed	0.990	0.553	0.709
MUS-appass_1_SptkBGCl	Baseline	0.998	0.571	0.727
	Proposed	0.956	0.695	0.805
MUS-mz_311_1_StbgTGd2	Baseline	0.997	0.590	0.741
	Proposed	0.992	0.681	0.808

5. Conclusion

In this paper, we propose a novel method for the onset detection of multiple-pitch in music signal. It is based on the observation that the starting points of the amplitude envelope always coincide with the onsets of some music notes in polyphonic music signal. Meanwhile, the temporal segmentation of music can effectively reduce the system pressure caused by time-frequency analysis at whole-file level. More details of each onset are obtained by segmental analysis, which leads to higher onsets recall than conventional methods. In temporal segmentation, a self-adaptive matched filter is used to detect the obvious onsets of amplitude envelope. Then based on these onsets, we divide the input signal into small segments. In the segmental onset detection process, we firstly use the RTFI to get the time-frequency representation, which is then converted into PES by summarizing the weighted harmonic partials. Then a simple peak picking method is used to detect onsets in each

segment. Finally, the onsets in all segments are combined with the segments' boundaries to yield the onset distribution of the whole file.

The proposed algorithm only uses a simple segmental onset detection method. In fact, other complex methods that give more accuracy results can also be applied to this framework. The music note onsets detected can be further used for higher-level music analysis tasks, just as automatic music transcription, content-based music retrieval, etc.

REFERENCES

[1] P. Masri and A. Bateman, "Improved modelling of attack transients in music analysis-resynthesis," in *Proceedings of the International Computer Music Conference*. Citeseer, 1996, pp. 100–103.

[2] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *Signal Processing Letters, IEEE*, vol. 11, no. 6, pp. 553–556, 2004.

[3] R. Zhou, M. Mattavelli, and G. Zoia, "Music onset detection based on resonator time frequency image," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1685–1695, 2008.

[4] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 6. IEEE, 1999, pp. 3089–3092.

[5] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1035–1047, 2005.

[6] G. Turin, "An introduction to matched filters," *Information Theory, IRE Transactions on*, vol. 6, no. 3, pp. 311–329, 1960.

[7] S. Pauws, "Cubyhum: a fully operational" query by humming" system." in *ISMIR*. Citeseer, 2002.

[8] Y. L. Wan, Z. G. Wu, R. H. Zhou, and Y. H. Yan, "Automatic transcription of piano music using audio-vision fusion," *Applied Mechanics and Materials*, vol. 333, pp. 742–748, 2013.

[9] J.-J. Ding, C.-J. Tseng, C.-M. Hu, and T. Hsien, "Improved onset detection algorithm based on fractional power envelope match filter," in *19th European Signal Processing Conference (EUSIPCO 2011) Barcelona, Spain*, 2011.

[10] R. Zhou, "Feature extraction of musical content for automatic music transcription," 2006.

[11] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 37–40.

[12] L. Song, M. Li, and Y. Yan, "Pitch estimation based on harmonic salience," *Electronics Letters*, vol. 49, no. 23, pp. 1491–1492, 2013.

[13] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1643–1654, 2010.