ONSET DETECTION IN PITCHED NON-PERCUSSIVE MUSIC USING WARPING-COMPENSATED CORRELATION

Olaf Schleusing, Bingjun Zhang, Ye Wang

School of Computing, National University of Singapore
audio@schleusing.de, {bingjun, wangye}@comp.nus.edu.sg

ABSTRACT

Automatically extracting temporal information from musical recordings is inarguably one of the most critical subtasks of many music information retrieval systems. In this paper we present a system for automatic note onset detection in pitched non-percussive (PNP) musical sounds, which is the most challenging audio signal group for this task. We propose a new approach based on stable pitch cues and signal energy. A computationally inexpensive method for feature extraction, which efficiently suppresses vibrato, is combined with information derived from the signal energy in the feature space. Onsets are localized by a median filter based peak picking method. The proposed method is tested against a database of annotated violin recordings, covering a wide range of tempo and playing styles like vibrato and staccato. Our system outperforms prior state of the art systems with results for True Positives of 91.2% and False Positives of 9.2%.

Index Terms— Music, Information Retrieval, Feature Extraction

1. INTRODUCTION AND RELATED WORK

Music is a highly structured and layered ensemble of sounds. On one level of abstraction, note events produced by the actions of an artist provide key information, which can serve as a baseline for extracting higher level musical information in many other Music Information Retrieval (MIR) tasks. Each note event comprises a specific note onset time, a duration and a pitch. Note onsets technically represent a transient segment of the audio signal, or more specifically they mark the instant in time, when the signal starts to evolve from a steady state to another steady state, i.e., from one note to the next. The detection of onsets, which is to localize the instants in time of these transitions, is what we are concerned with in this paper.

Most traditional approaches to onset detection [1] perform an optional preprocessing step and then derive an intermediate signal or detection function, which is at a significantly lower sampling rate than the audio signal itself and reduces the signal to a more favorable representation. The key characteristic of this representation or detection function is a well suited feature, which highly correlates with the musical cues it intends to unveil. This detection function can then be fed into a peak picking method to find local maxima, i.e. onsets. Recently more sophisticated approaches have been presented to comprise higher level musical information via machine learning or statistical methods for extracting the musical cues from various intermediate representations (e.g., [2]). In this paper we focus on feature extraction while employing an existing simple peak picking method.

A comprehensive evaluation of different detection functions and their suitability for different types of audio signals was published recently in [1]. The authors outlined and compared different methods for finding onsets in musical signals and provided information on which methods work best for which signal class and which application. Based on these comparisons Collins set up a more comprehensive evaluation of onset detection functions, which is described in [3]. Subsequently he proposed a new method specifically designed for PNP sounds [4]. This method was developed under the assumption that the perception of stable pitch cues could be linked to the segmentation of notes. He employed a pitch tracker and derived onset information directly from changes in pitch and signal energy.

In this paper we present a new method, also targeted at PNP sound. In particular we focus on violin music produced from a single instrument and recorded in a home environment. Due to room reverberations and resonances of the violin body, the decaying harmonics of a note often overlap with the harmonics of the next note, especially when legato is played. In addition, string instruments also provide the possibility to excite more than one string of the instrument at the same time, i.e. double stops or triple stops. This and the non-percussive playing style of bowed string instruments like the violin make note onsets particularly difficult to locate. Most difficult cases are represented by consecutive notes, where there is only a small dip in loudness while the pitch is unchanged.

The rationale behind the method presented in this paper is based on the observation, that for PNP musical sound, as it is produced by stringed, bowed instruments, the most reliable cue to locate onsets are spectral fluctuations. Our method calculates an average of spectral change in consecutive frames to generate a detection function and is inspired by an algorithm originally presented by Boo et al [5].

In this paper we first outline the algorithm used to derive the onset locations in section 2. Section 3 describes the database we used for evaluating our system and outlines a recent method, which we re-implemented for comparison and evaluation. In section 4 we present the results of this evaluation, followed by discussions and possible future work in section 5.

2. PROPOSED METHOD

Our system follows a typical framework by first deriving a sub-sampled detection function from the audio signal itself. A peak picking method is then used to locate the onsets from this detection function.

2.1. Detection Function

The key is to find features, which highly correlate with musical onsets, for the translation of a musical signal into a detection function. For PNP sounds it is characteristic, that resonances are formed within the body of the instrument very quickly after the string is excited. These resonances are relatively stable while the note is played at a constant pitch without vibrato. To exploit this observation, we employed the correlation between time-wise consecutive frames of the audio signal to localize regions of stable pitch. The Short-Time Fourier Transform (STFT) with a window length of 1024 samples with a sampling rate of 44.1 kHz and 50% window overlap was used to transform the audio signal into the frequency domain. The fact that the energy of higher order harmonics is decreasing with a rising number of the order led to the exclusion of higher frequencies from the correlation calculation. The explanation for that is, that the signal-to-noise ratio is bound to become smaller with rising frequencies due to a smaller signal energy. According to [1] and our observations, the information in higher frequency bands, as exploited by the HFC method [6] is useful for feature extraction from percussive music, but not from non-percussive signals. Higher frequency bands usually lead to a greater contribution of noise to the resulting correlation, which is undesirable since it lowers the total correlation. Experiments showed, that a cutoff at 8 kHz yielded the best results. Due to the same observation these lower frequencies were split into three subbands of equal bandwidths. Each of these bands received a specific weight to account for the difference in signal energy.

The detection function is then formed by the following equation:

$$d(t_0) = 1 \Bigg/ \sum_{f=t_0-\lfloor \frac{w}{2} \rfloor}^{t_0+\lfloor \frac{w}{2} \rfloor} \prod_{b=1}^{b_0} \left(\left(\frac{c_b(t_0, f)}{\sqrt{c_b(t_0, t_0) c_b(f, f)}} \right)^{W_b} \right) \tag{1}$$

where w is the number of frames used to calculate and average the correlation, t_0 is the frame number of the spectrogram, c_b is the covariance of the b-th subband of two frames and W_b is the weighting factor for that subband. b_0 is the number of subbands in one frame. The product leads to an amplification of the detection function upon onsets in any of the three subbands.

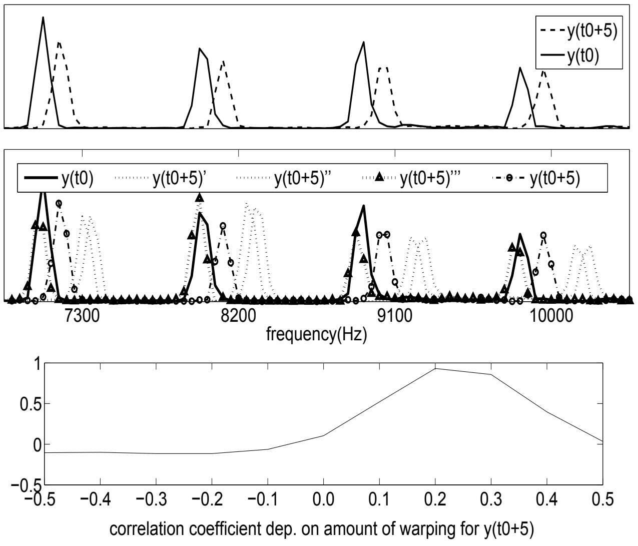

Fig. 1. Warping compensation: In the upper figure fractions of two frames y of an audio signal during vibrato performance are displayed, 60 ms (5 frames) apart from each other. Clearly the shift of the frequency due to the vibrato can be seen; the harmonics of $y(t_0 + 5)$ emerge at slightly higher frequencies. That yields a low correlation (e.g. a correlation coefficient of 0.104 in band 3) between $y(t_0)$ and $y(t_0 + 5)$. In the middle figure y(t0), y(t0+5) and a selection of the 10 warped signals of y(t0+5) is drawn, with y(t0+5)''' being warped by 0.2 times a semitone yielding the best match and thus also the best correlation coefficient (0.932). The lowest graph shows the correlation coefficient between $y(t_0)$ and $y(t_0 + 5)$ for $y(t_0 + 5)$ being warped from minus to plus half a semitone.

The resulting detection function sports values close to zero during periods of stable pitch and comprises peaks on occasions where the spectral characteristic of the audio signal changes. As formulated in Equ. (1), each sample of the detection function is calculated from correlation coefficients of w neighboring frames. The reason is, that considering more than just one pair of frames averages out contributions through randomly similar frames and flattens the detection function during periods of stable pitch. An upper bound to w is determined by the music characteristics. In our case we used 13 neighboring frames, which corresponds to around 150 ms. From a musical point of view, this is a period that is shorter than a quaver (for a tempo slower than 200 BPM) and thus should guarantee stable pitches for this period in most cases.

When it comes to musical background, one can easily spot that slight changes in pitch would disturb the reliabil-

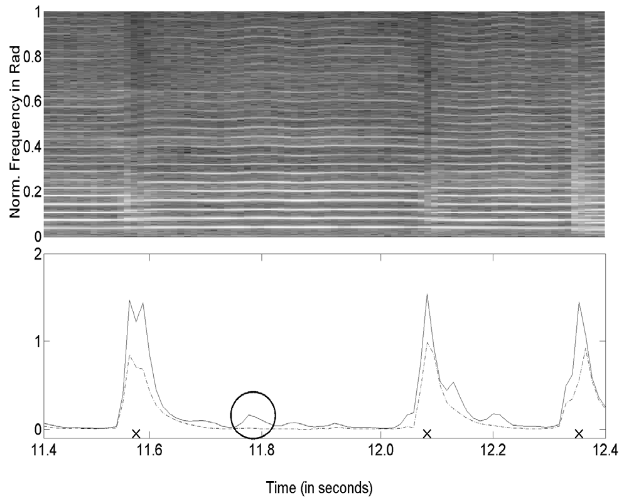

Fig. 2. The effect of warping on the detection function. The upper figure displays a 1 second section of the spectrogram of an audio signal with vibrato. In the lower figure, the locations of the onsets are denoted with crosses and two detection functions are drawn. The solid line represents a detection function without warping correction, the dashed line is corrected. Clearly can be seen (as highlighted in the circle) that spikes in the detection function caused by vibrato are suppressed by the correction. Another effect though is, that the amplitude of the detection function at onset locations is lowered

ity of correlation. Vibrato, an expressive feature widely used among musicians playing stringed, bowed instruments; is a challenge to any onset detection method aiming PNP sounds. Vibrato is produced by rolling the stopping finger on the string up and down, effectively shortening and lengthening the excited string. This leads to oscillating changes in amplitude and pitch of usually less than +/- half a semitone. Although the time-wise consecutive frames of slightly different pitches appear to be very similar, the inverse correlation is sensitive to even the slight changes in pitch. To account for this we introduced a new method called warp-compensation. Equ. (1) is modified in that each correlation coefficient c_b not only represents the covariance of the frames at frame t_0 and f, but is the maximum of the covariance of these two frames, where the spectrum at f is warped in 10 equidistant steps from minus half a semitone up to plus half a semitone (Fig. 1). This efficiently compensates for small pitch changes due to vibrato. A comparison between detection functions with and without warping is illustrated in Fig. 2.

2.2. Peak Picking

Before the detection function is analyzed by a peak picking method, it undergoes a preprocessing step. Due to the nature of the detection function generation, the signal energy is completely uncorrelated to the detection function. That means, not only changes in the spectrum cause peaks in the detection function, but also very quiet portions of the signal may do so due to noise. This is particularly severe when a piece is performed in staccato playing style where each note is cut short and followed by a very short period of silence. For this reason the signal energy is used to mute sections of the detection function, where the musical signal is absent or of very low energy.

Finding candidates of onsets from the detection function comprises the localization of local maxima in the detection function. This can be done by simply searching for peaks, steep rising edges or some other characteristic shapes, depending on the feature used to create the detection function. Our method searches for peaks after applying a threshold, which is adapted to temporal fluctuations of the signal with the help of a median filter. This method is particularly useful for reducing the influence of occasional high spikes of the detection function on the adaptive threshold and is similar to the method presented in [1]:

$$\tilde{\delta}(n) = \delta + \lambda median\{|d(n-M)|, ..., |d(n+M)|\} \quad (2)$$

where $\tilde{\delta}(n)$ is the adaptive threshold as a function of n, δ is a base threshold, λ is a weighting constant, $d(n)$ is the detection function and M is the window length of the median filter. All peaks above this threshold are then combined to a single peak if they happen to be not further apart than 50 ms from each other and then this peak is marked as an onset.

3. EVALUATION

In this section we give a short description of the database used for the evaluation and describe how we evaluated the performance of the onset detection system.

3.1. Database description

All pieces used throughout the evaluation were recorded in an ordinary room equipped with sound absorbing curtains to reduce reverberations. The pieces were selected to cover a wide range of violin performances of different playing styles and tempo ranging from slow to fast. 50% of the pieces contained vibrato, several were played in staccato style and some contained double stops (multiple strings played simultaneously), slurs and legato. Overall the database contained 9717 onsets and was annotated by subjects with musical education. For the annotation process itself a proprietary tool was used, developed specifically for onset annotation.

3.2. Evaluation method and results

In order to evaluate the performance of the presented method we re-implemented a method presented by Collins in [4], which

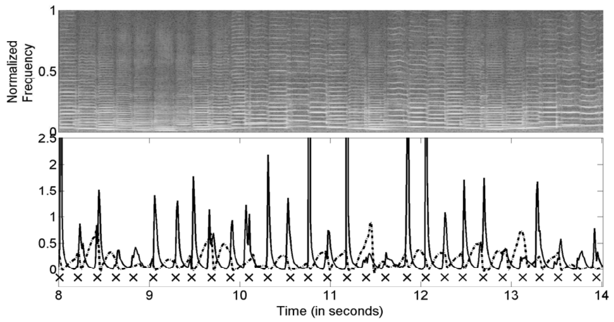

Fig. 3. Comparison of detection functions. The upper figure displays a portion of the spectrogram of a performance of Franz Wohlfahrt's Etudes Op.45/34. In the lower figure both detection functions are drawn as well as the humanly annotated note onsets (as crosses). As can be seen clearly, the inverse correlation method (solid line) sports clear peaks around note onsets while featuring a low signal during notes being played. On the other hand, the peaks in Collins' method (dashed line) are less sharp and its performance suffers from overlapping notes (legato) due to its dependency on pitch contours.

was also developed to address onset detection in PNP music. In our re-implementation we used the YIN-algorithm [7] as a front end for the extraction of pitch contours and achieved similar results as originally reported.

As a measure for detection performance we employed the widely used notation of *True Positives (TP)* and *False Positives (FP)*, where the definition is as follows:

$$TP = \frac{NumberofCorrectlyFoundOnsets}{NumberofallFoundOnsets} \tag{3}$$

$$FP = \frac{NumberofCorrectlyFoundOnsets}{NumberofallAnnotatedOnsets} \tag{4}$$

For evaluating correctly extracted onsets a tolerance of 70 ms was used. Overall the reference implementation from [4] yielded a TP of 62.37% and a FP of 24.43%. The presented warp-compensated inverse correlation achieved a TP of 91.2% and a FP of 9.2%

4. DISCUSSION AND FUTURE WORK

In this paper we presented a new method for feature extraction from PNP musical sounds for onset detection, which clearly outperforms existing state of the art methods for this task. The algorithm accurately discriminates onsets from non-onsets for this signal class and yielded excellent results when tested against our database in comparison to the state-of-the-art for PNP sounds. Several problems of the current implementation like the vulnerability of the detection performance to the presence of staccato could be solved by incorporating higher level, musical knowledge, i.e. in a Bayesian framework or in a framework recently presented by Klapuri et al. in [2]. Another solution could be to dynamically adapt the width and number of bands considered for the correlation. When staccato is performed, the signal in higher frequency bands decays earlier and much faster than in lower bands and thus emphasizes the undesirable effect of noise in higher bands on the detection function. Nevertheless we have shown that the inverse correlation method is capable of very accurately unveiling temporal information for PNP sound and is a promising candidate for use as a front-end in Music Information Retrieval tasks.

5. REFERENCES

[1] J. B. Bello, L. Daudet, A. Samer, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. on Speech and Audio Processing*, pp. 1035– 1047, 2005.

[2] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 1, pp. 342–355, January 2006.

[3] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psycho-acoustically motivated detection functions," in *Convention Preprint of 118th AES Convention*, Barcelona, 2005, Preprint No. 6363.

[4] N. Collins, "Using a pitch detector for onset detection," in *Proc. of ISMIR2005*, Barcelona, Spain, 2005, pp. 100–106.

[5] J. Boo, Y. Wang, and A. Loscos, "A violin music transcriber for personalized learning," in *ICME*, Toronto, Ca, 2006, pp. 2081–2084.

[6] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*, Ph.D. thesis, University of Bristol, UK, 1996.

[7] A. Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoust. Soc. of America*, vol. 111, no. 4, pp. 1917–1930, April 2002.