A REASSIGNED BASED SINGING VOICE PITCH CONTOUR EXTRACTION METHOD

Georgina Tryfou[1,2] *and Maurizio Omologo*[1]

[1]Fondazione Bruno Kessler (FBK)
[2]University of Trento - Unitn

ABSTRACT

Although there are many systems concerned with melody extraction from polyphonic music, there are certain limitations stemming from the spectral processing that are yet to be overpassed. In this paper, we propose a novel method to create sets of melodic pitch contours which are shown to contain harmonic information critical for a melody extraction system. The proposed approach exploits interesting characteristics of the reassigned spectrogram and computes a new representation which comprises a set of points in the time-frequency domain, weighted according to their dominance, in terms of harmonic content. The experimental results show that the proposed method is a valid approach to the detection of time-frequency points that are related to the melodic content of music signals. Moreover, the quality of the acquired melodic pitch contours is proved through a comparison with those extracted by a state-of-the-art melody extraction system.

Index Terms— Melody extraction, reassigned spectrogram, singing voice, pitch contours

1. INTRODUCTION

The increasing interest in music related applications, for example the automatic transcription of audio recordings, the creation of karaoke files, and the music retrieval by singing or humming, has recently led to extensive research activities in the area of modelling the vocal melody of real world music recordings[1, 2]. According to the prerequisites of each specific application, the melody line has to be described in terms of a sequence of frequencies, transcribed into sung or played notes, or expressed in terms of vocal effects, as for example tremolo and vibrato.

The task of melody extraction is closely related to pitch detection and, due to the similarities of the two tasks, the first successful solutions were inspired by the extensive literature in the area of pitch extraction. However, the nature of music signals brings limitations in the success of such methods. First, a music signal may comprise many different instruments, with two or more notes from the same, or different, instruments sounding simultaneously. Furthermore, percussive sounds and inharmonicities may, in principle, take place at any moment and the vocal melody may interfere with partials of different sounds. For all these reasons, the use of pitch extraction methods, that are designed for speech or monophonic sound, does not produce the necessary results when singing voice, or another melody source in the context of polyphonic music is considered. On the other hand, current melody extraction methods can face problems when, for instance, the target harmonic content is vocal. The fine structure of a vocal melody during vocal effects such as tremolo and vibrato can be hardly characterized with the commonly used spectral representations, such as the short time Fourier transform (STFT).

In this work we propose a novel method to extract pitch contours that describe the melodic content of music signals. The strength of this system lies in the proposed spectral representation, called dominance reassigned spectrogram (DRS). Since this representation utilizes the reassigned spectrogram (RS) it offers a much better localization of the energy of the various harmonic components. In addition, the dominance weighting that we propose exploits the characteristics of the RS in order to add extra salience to the components that are related to the predominant melody. Furthermore, we describe a way to map the sporadic data of the DRS into melodic pitch contour (MPC) sets.

The remainder of this paper is organised as follows. In Section 2 we review the existing work in the area of melody line extraction, with a focus on the spectral processing commonly applied, and introduce the RS. In Section 3 the details of the proposed method are described. The experimental activities and related results are presented in Section 4. Finally, in Section 5 we conclude this paper with an overview of the proposed method and the results.

2. RELATED WORK

Based on the approach selected to process the input signal in the spectral domain, melody extraction systems can be divided into two broad categories: the *salience-based* and the *separation-based*. Methods in the first category transform the input audio signal into a pitch salience signal, where each frequency is associated with a certain value of energy or salience. Sub-harmonic summation (SHS) is very commonly used in order to create the salience signal [3, 4, 5]. On the

other hand, *separation-based* approaches [6, 7] attempt to segregate the singing voice from the music accompaniment and perform the melody line detection on the segregated vocal signal. *Hybrid approaches*, where spectral processing is based on a salience signal extracted after an initial harmonic/percussive separation step, were proposed in [8, 9].

After the spectral processing, that commonly results in a multi-pitch representation of the input music signal, the melody line is tracked. This tracking is done through dynamic programming [8, 10, 7], tracking agents [11, 12], or hidden Markov models (HMM) [13, 14, 9]. Finally, a voicing detection part, often incorporating some type of thresholding [5, 15, 16], determines in which regions in time the melody line is active. In this work, we focus on the first part of melody extraction systems, *i. e.,* the spectral and multi-pitch representation, and do not address the melody line tracking and the voicing detection parts.

2.1. Spectral Representations

The vast majority of melody extraction systems exploits the STFT for the transformation of the input signal into the spectral domain. The problems that arise from the use of the STFT in the context of many diverse signal processing applications are extensively discussed in the literature. These concern the unavoidable trade-off between the time and frequency resolution, and the fact that the selected resolution is fixed over the whole spectrum. However, in melody detection different frequency resolutions in the various spectral regions can be highly beneficial. For these reasons, alternative representations were proposed for the tasks of pitch tracking and melody extraction, as for example Multi-Resolution FFT [17, 8, 9], multirate filterbanks [12] and constant-Q transform [15]. Other approaches include frequency and time correction mechanisms as the ones discussed in [18], parabolic interpolation as in [19] and instantaneous frequency (IF) calculation as in [17, 4].

2.2. The reassigned spectrogram

The RS, firstly introduced in [20], provides an estimation of the IF, by assigning the spectral energy of each analysis frame closer to its true region of support. It has been successfully used in the past for other signal processing tasks, for example voice identification, phonation analysis, and visualization of the formant structure in speech [21], automatic chord recognition [22] and automatic speech segmentation [23]. Although a successful method to separate the components from impulses in the RS has been thoroughly described in [24], and the use of RS for improved estimation of IF has been successfully applied in [25], there is no prior work that utilizes the RS as the spectral representation of music signals for melody tracking. In [26], the RS of musical audio data was analysed in an early attempt to use it within a front-end for transcription, but no further work supported this initial study.

The mathematical formulation of the RS begins with the STFT of a signal, $X(t,\omega)$. The energy of the point (t,ω), where t is the time frame and ω is the frequency bin, is reassigned at a new point that better reflects the distribution of energy of the analysed signal. The time-frequency reassigned (TFR) coordinates $(\hat{t},\hat{\omega})$ are calculated from the derivatives of the spectral phase $\phi(t,\omega)$ as follows

$$\hat{t}(t,\omega) = -\frac{\partial\phi(t,\omega)}{\partial\omega} \quad (1)$$

$$\hat{\omega}(t,\omega) = \omega + \frac{\partial\phi(t,\omega)}{\partial t} \quad . \quad (2)$$

In practise, spectral energy from the coordinate (t,ω) is reallocated to coordinate $(\hat{t},\hat{\omega})$, the latter one defined in the continuous time and frequency domains.

3. PROPOSED METHOD

The proposed processing starts with a preprocessing step, where an equal loudness filter is applied in order to enhance the frequencies where the melody line is normally found. The next steps, *i.e.,* DRS representation and MPC extraction are introduced in the following.

3.1. DRS representation

As described in [27, 28], the STFT points that correspond to a minimum distance between the IF and the center of the spectral bins are strong indicators of the presence of fundamental frequencies in this spectral region. The above observation can be extended to the RS, where a minimum frequency reassignment is observed in the regions of harmonic components. Therefore, the dominance RS (DRS) is defined as

$$D(\hat{t},\hat{\omega}) = \left(\frac{X(\hat{t},\hat{\omega})}{\omega-\hat{\omega}}\right)^2 \quad , \quad (3)$$

where $X(\hat{t},\hat{\omega})$ is the power RS and ω denotes the frequency from which the reallocation originated. The difference $\omega-\hat{\omega}$, *i. e.,* the amount of frequency reassignment from the spectral point (t,ω) to the corresponding TFR point, is minimized in the region around dominant components, leading to a maximization of $D(\hat{t},\hat{\omega})$. Further salience is added to the most dominant TFR points with the use of the power law. In practice, the DRS assigns a degree of dominance to each TFR point, describing the importance of this point in terms of melodic content. This novel spectral representation of the music signal is used in the subsequent processing and enables the proposed multi-pitch extraction that results is sets of melodic pitch contours (MPC).

3.2. MPC extraction

Here, we propose a method that selects TFR points that are related to melodic content and groups them into MPC. The goal

is to detect regions of the DRS where the TFR points are *connected*, and then, assign these points to unique pitch contours. The detected *connectivity* is a strong indication that in the corresponding RS region there is an underlying structure, which is related to the melodic components. In order to create the pitch contours we built upon the fact that more TFR points, of higher dominance, are found around the melodic components. The iterative Algorithm 1 is proposed as a means to determine the set of MPC.

Algorithm 1 MPC extraction

1: **Input:** The dominance RS $D(\hat{t},\hat{\omega})$, the RS $X(\hat{t},\hat{\omega})$.
2: $E_{total} \leftarrow \sum_{\forall(\hat{t},\hat{\omega})} X(\hat{t},\hat{\omega}), \quad E_{contours} \leftarrow 0$
3: **while** $E_{contours} \leq r_{min} E_{total}$ **do**
4: Initialize a new pitch contour, C
5: $N_0 \leftarrow \arg\max_N \sum_{(\hat{t}_n,\hat{\omega}_n)\in N} D(\hat{t}_n,\hat{\omega}_n)$
6: $i \leftarrow 0$
7: **while** $|N_i| < N_{min}$ **do**
8: $P_c(N_i) \leftarrow$ centerOfGravity(N_i)
9: Add $P_c(N_i)$ in C
10: Remove $P_c(N_i)$ from $D(\hat{t},\hat{\omega})$
11: $N_{i+1} \leftarrow$ getNeighbourhood$(P_c(N_i))$
12: $i \leftarrow i+1$
13: **end while**
14: $E_{contours} \leftarrow E_{contours} + \sum_{(\hat{t},\hat{\omega})\in C} X(\hat{t},\hat{\omega})$
15: **end while**

A neighbourhood N, of a central TFR point $(\hat{t}_c,\hat{\omega}_c)$ is defined as the spectral area that contains all the spectral points for which $|\hat{t}_c - \hat{t}| \leq \Delta\hat{t}$ and $|\hat{\omega}_c - \hat{\omega}| \leq \Delta\hat{\omega}$, where $\Delta\hat{t}$ denotes the maximum allowed time deviation from the center of the neighbourhood and $\Delta\hat{\omega}$ the maximum allowed frequency deviation. On the other hand, given a neighbourhood N, the central TFR point $(\hat{t}_c,\hat{\omega}_c)$ can be found as the *center of gravity* of it, as follows

$$(\hat{t}_c,\hat{\omega}_c) = \frac{1}{D_N}\left(\sum_{(\hat{t}_n,\hat{\omega})\in N} D(\hat{t}_n,\hat{\omega})\hat{t}_n, \sum_{(\hat{t},\hat{\omega}_k)\in N} D(\hat{t},\hat{\omega}_k)\hat{\omega}_k\right), \tag{4}$$

where D_N is the local dominance of N, calculated as

$$D_N = \sum_{(\hat{t}_n,\hat{\omega}_n)\in N} D(\hat{t}_n,\hat{\omega}_n) \quad . \tag{5}$$

At each outer iteration of the algorithm 1, the neighbourhood with the highest local dominance is selected as the starting point of a new pitch contour (see line 5). In the inner iteration, the *center of gravity* $P_c(N_i)$ of the neighbourhood under consideration is added to the current contour. The same point is used in order to update the neighbourhood before the following iteration, as described above, and then it is removed from the DRS. The contour tracking continues with the remaining points and it is exhaustive, meaning that a contour ends when the newest created neighbourhood is empty, or its cardinality $|N_i|$ reaches a certain threshold N_{min}, and both directions in time have been checked. The outer iteration stops when the energy of the created contours, $E_{contours}$, is more than a certain ratio, r_{min}, of the total energy, E_{total}, of the musical excerpt. The selection of N_{min} has been experimentally defined to 15 TFR points. The setting of r_{min} is discussed in Section 4.

After the extraction of the MPC, a post-processing step that detects and corrects harmonic sets is applied. The processing is based on the SHS matching theory of [29], which inspired a very successful pair wise evaluation of spectral peaks, proposed in [30]. Here, we use the same idea of pairwise comparison of pitch contours in order to detect harmonic sets and correct them by adding missing harmonic roots.

4. EXPERIMENTS AND RESULTS

4.1. TFR point-wise evaluation

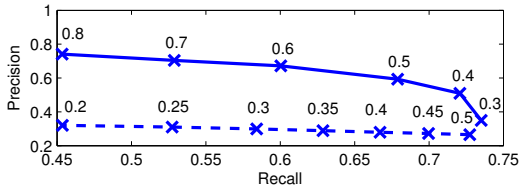
Here, we study the ability of the proposed method to correctly identify the set of TFR points that are related to the melodic content of the piece. We compare the behaviour of the proposed method to a baseline method, which comprises imposing the mixed partial derivative (MPD) criterion of [24]. According to this approach, the set of points that are related to the harmonic content of the signal are those that meet the following condition

$$\left|\frac{\partial\phi^2(t,\omega)}{\partial t\partial\omega}\right| < A \quad , \tag{6}$$

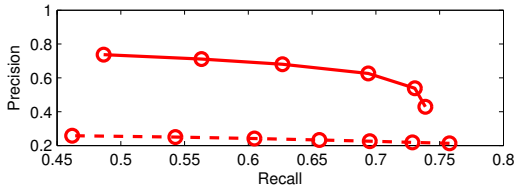
where A is a tolerance factor that defines the maximum variation of an accepted component from the ideal sinusoid. In order to quantify the results, we define the point precision/recall and f-measure metrics, which are calculated over the number of TFR points that are retrieved. A retrieved point is considered relevant if it lies within half semitone from the annotated melody line.

There is a trade-off between how precise (point precision) and how sensitive (point recall) each TFR point selection method is. In the proposed method, the exact behaviour in terms of precision/recall is controlled by the parameter r_{min} of Algorithm 1. In the MPD method this is controlled by the threshold value A introduced earlier. In Figure 1, the point precision/recall curves for two different datasets are presented. Experiments were conducted using data sampled at $44.1kHz$, with a window size of $30ms$ and a step of $5ms$. The MPC extraction is based on neighbourhoods created with $\Delta\hat{t}$ of $15ms$ and $\Delta\hat{\omega}$ of 0.5 semitones.

Although different parameters are set for the proposed and the baseline methods in order to produce these curves, the comparison is meaningful, since, in practice, each of the parameters designates the strictness of the point selection method. From the curves, it is evident that the proposed

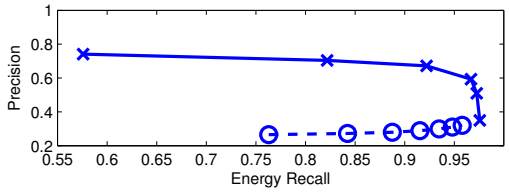


(a) ADC2004

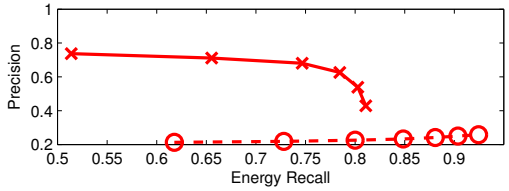


(b) Mirex05

Fig. 1: The Precision/Recall curves of the proposed (solid) and the MPD (dashed) methods in creating sets of TFR points related to the melodic content. As shown in the top figure, the points correspond to different r_{min} values for the proposed method, in the range 0.3 to 0.8, and different A values for the MPD method, in the range 0.5 to 0.2. The order of the points is the same in the subsequent figures.



(a) ADC2004



(b) Mirex05

Fig. 2: The precision/energy recall curves of the proposed (solid) and MPD (dashed) methods in creating sets of TFR points related to the melodic content. The points correspond to different r_{min} and A values for the MPD method, as described in earlier experiment.

method is much more precise that the baseline in selecting TFR points that are related to the melody line.

In addition, in order to study the importance of the missed TFR points, we define the energy recall metric, which is the spectral energy sum for all the melody points tracked by the algorithm, divided by the total energy of the signal. In Figure 2, the point precision of the proposed and MPD methods is depicted, as a function of the energy recall of each method. We observe that both methods are successful in selecting the TFR points that bare the most significant amount of energy of the harmonic components. Furthermore, it is shown that the MPD method is able to produce higher energy recall measures, especially in the case of Mirex05. Nevertheless, the corresponding precision values are too low to yield any useful application.

4.2. Evaluation of the MPC

The post-processed MPC are mapped in a grid as in [23] and are evaluated with the contour precision/recall and f-measure metrics, as in [31]. The same evaluation is applied on the contours extracted with the MELODIA [1] vamp plug-in. The comparative results are presented in Table 1. As shown there, in both test datasets the proposed contour extraction method results in a higher f-measure, than the MELODIA. Particularly, for the ADC2004 dataset the proposed method improves both the precision and recall metrics. Although this is not the same for the Mirex05 dataset, the proposed method is producing more balanced precision/recall pairs of values, and therefore results in higher f-measures for both datasets. This is an interesting finding as it means that the selection process that leads to the pitch contours, *i. e.,* the DRS and the multi-pitch extraction algorithm, is more successful than the literature method in retrieving points that are actually related to the melody line.

[1] http://mtg.upf.edu/technologies/melodia

Dataset	Method	Pr	Re	F
ADC2004	MEL.	0.58	0.7	0.63
	Pr.	0.75	0.8	0.76
Mirex05	MEL.	0.48	0.77	0.59
	Pr.	0.48	0.73	0.63

Table 1: The contours formed by the MELODIA vamp plugin (MEL.) are compared to those extracted by the proposed method (Pr.). The average f-measure F is computed over the f-measure of all the excerpts.

5. CONCLUSIONS

In this paper, we presented a method that detects the spectral regions of polyphonic music signals where melodic components are active, and groups these components in harmonic sets of MPC. The use of the RS in the core of the system provides a set of finely tuned contours that ensure the minimization of errors related to the limitations of the STFT. The MPC extraction algorithm is based on a dominance weighting of the TFR data. The experimental activities showed that the proposed method is superior to the MPD criterion in selecting points related to the harmonic content of the signal. Furthermore, the produced pitch contours scored a higher f-measure than a state-of-the-art system. As a next step, we are interested to incorporate to the proposed method the tracking and voicing detection modules, in order to extract highly accurate melody lines from polyphonic music signals.

6. REFERENCES

[1] Anssi Klapuri and Manuel Davy, *Signal processing methods for music transcription*, Springer, 2006.

[2] Markus Schedl, Emilia Gómez, and Julián Urbano, "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.

[3] Matti Ryynänen, *Automatic Transcription of Pitch Content in Music and Selected Applications*, Ph.D. thesis, Tampere University of Technology, 2008.

[4] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[5] K. Dressler, "Audio melody extraction for mirex 2009," in *Music Information Retrieval Evaluation eXchange*, 2009, MIREX.

[6] Jean-Louis Durrieu, Gael Richard, and Bertrand David, "An iterative approach to monaural musical mixture de-soloing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, ICASSP, pp. 105–108.

[7] Hideyuki Tachibana, Takuma Ono, Nobutaka Ono, and Shigeki Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, ICASSP, pp. 425–428.

[8] Chao-Ling Hsu and J.-S.R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[9] Tzu-Chun Yeh, Ming-Ju Wu, J.R. Jang, Wei-Lun Chang, and I-Bin Liao, "A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, ICASSP, pp. 457–460.

[10] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2145–2154, 2010.

[11] Masataka Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[12] Masataka Goto, "PreFEst: A predominant-f0 estimation method for polyphonic musical audio signals," in *Music Information Retrieval Evaluation eXchange*, 2005, MIREX.

[13] Matti Ryynänen and Anssi Klapuri, "Transcription of the singing melody in polyphonic music," in *7th International Society for Music Information Retrieval Conference*, 2006, ISMIR.

[14] Christopher Sutton, Emmanuel Vincent, D. Plumbley, Mark, P. Bello, Juan, Christopher Sutton, Emmanuel Vincent, D. Plumbley, Mark, and P. Bello, Juan, "Transcription of vocal melodies using voice characteristics and algorithm fusion," in *Music Information Retrieval Evaluation eXchange*, 2006, MIREX.

[15] Pablo Cancela, "Tracking melody in polyphonic audio. Mirex 2008," in *Music Information Retrieval Evaluation eXchange*, 2008, MIREX.

[16] Graham E. Poliner and Daniel P. W. Ellis, "A classification approach to melody trascription," in *6th International Society for Music Information Retrieval Conference*, 2005, ISMIR, pp. 161–166.

[17] Karin Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution fft," in *9th International Conference on Digital Audio Effects*, 2006, DAFx, pp. 247–252.

[18] Florian Keiler and Sylvain Marchand, "Survey on extraction of sinusoids in stationary sounds," in *5th International Conference on Digital Audio Effects*, 2002, DAFx, pp. 51–58.

[19] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[20] Kunihikq Kodera, Roger Gendrin, and Claude de Villedary, "Analysis of time-varying signals with small BT values," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 64–76, 1978.

[21] Sean A Fulop, *Speech Spectrum Analysis*, Springer Berlin Heidelberg, 2011.

[22] Maksim Khadkevich and Maurizio Omologo, "Reassigned spectrum-based feature extraction for GMM-based automatic chord recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–12, 2013.

[23] Georgina Tryfou, Marco Pellin, and Maurizio Omologo, "Time-frequency reassigned cepstral coefficients for phone-level speech segmentation," in *22nd IEEE European Signal Processing Conference*, 2014, EUSIPCO, pp. 2060–2064.

[24] S. A. Fulop and K. Fitz, "Separation of components from impulses in reassigned spectrograms," *Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1510–1518, 2007.

[25] Masashi Ito and Masafumi Yano, "Sinusoidal modeling for nonstationary voiced speech based on a local vector transform," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1717–1727, 2007.

[26] Stephen W Hainsworth, Malcom D Macleod, and Patrick J Wolfe, "Analysis of reassigned spectrograms for musical transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 23–26.

[27] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity.," in *EuroSpeech*, 1999, vol. 99, pp. 2781–2784.

[28] Tomohiro Nakatani and Toshio Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3690–3700, 2004.

[29] Ernst Terhardt, Gerhard Stoll, and Manfred Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.

[30] Karin Dressler, "An auditory streaming approach for melody extraction from polyphonic music," in *12th International Society for Music Information Retrieval Conference*, 2011, ISMIR, pp. 19–24.

[31] J. Salamon, E. Gómez, and J Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," in *14th International Conference on Digital Audio Effects*, 2011, DAFx, pp. 73–80.