PREDOMINANT MELODY EXTRACTION FROM VOCAL POLYPHONIC MUSIC SIGNAL BY COMBINED SPECTRO-TEMPORAL METHOD

Gurunath Reddy M, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology, Kharagpur, India

{mgurunathreddy, ksrao}@sit.iitkgp.ernet.in

ABSTRACT

A combined spectro-temporal based method is proposed to derive the predominant melody from vocal polyphonic music signals. In the proposed method, vocal (voiced) and non-vocal (unvoiced) segments are determined by strength of excitation. The vocal segments are further divided into voiced notes by detecting their onsets using transition cues present in spectral domain. The melody contour present in each of the voiced note segments is obtained by using an adaptive zero frequency filtering (ZFF) in time domain. The process of melody extraction is provided in more detail and the initial results showed the potential use of the proposed method for vocal melody extraction.

Index Terms: Predominant Melody, Zero Frequency Filter, Note Onsets, Strength of Excitation, Note Boundaries.

1. INTRODUCTION

Predominant melody extraction is the task of automatically extracting the fundamental frequency (F0) contour of the dominant musical instrument in a polyphonic music signal. In a polyphonic music signal, the dominant musical instrument can be human singing voice or a lead instrument. The accurate extraction of melody has many potential applications [1] such as Query by Humming, singer identification, automatic music transcription, music genre classification, computational auditory scene analysis and many more. Since the human singing voice is dominant in most of the polyphonic music signals, vocal F0 contour extraction is the main objective of this paper.

In literature, we can find two major approaches to derive the melody of a music signal viz. *Source separation* and *Salience based* methods. Source separation based methods extract the F0 contour of the melody source by separating it from rest of the music signal by modelling melody source and the accompanied instruments separately [2, 3, 4, 5]. On the other hand, melody contour is extracted by estimating the pitch salience of the composite signal over time in salience based methods. Then, F0 tracking algorithms are applied on the estimated salience to obtain melody contour of the predominant melody source. Salience based methods mostly differ by the way the salience function is computed, salience peaks are estimated and the melody contours of the dominant source created by pitch tracking methods [6, 7, 8, 9, 10]. For a detailed review on salience and source separation based methods Ref. [1].

In this paper, we propose a filtering based method to extract vocal melody from polyphonic music signals. In this work, the bandpass filtering nature of the zero frequency filtering (ZFF) method is exploited to extract the instants of significant excitation (epochs) or Glottal closure instants (GCI) of the vocal melody source from mixture signal. Originally, ZFF is proposed to extract the F0 contour of the monophonic speech signal by filtering it with a zero frequency resonator (ZFR) followed by designing a narrow bandpass like filter in time domain with center of frequency equal to the average pitch period of the speaker. The same method cannot be applied for the music signal because of the composite nature of the signal consisting of many pitched sources. Also the pitch of the lead voice varies significantly from one note to the other. Hence, a single filter with center of frequency equal to the average pitch period is not sufficient to obtain the accurate F0 of the lead voice from the entire music signal. To overcome these limitations, the polyphonic music signal with lead voice is segmented into voiced note like regions by identifying the note onsets. For each identified note, a representative average pitch period is obtained by Two-Way Mismatch (TWM) algorithm. Then each note is filtered adaptively by designing time domain bandpass filter with center of frequency corresponding to the representative pitch period to obtain the melody F0. Initially, the voiced and unvoiced regions are obtained by thresholding the strength of excitation (SoE) contour obtained by the ZFF signal.

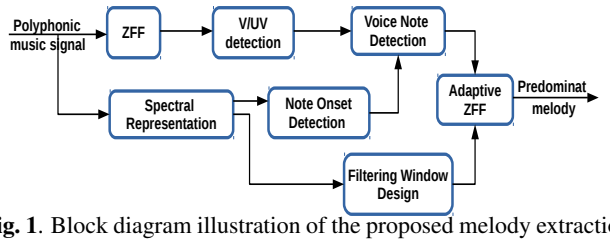

Fig. 1. Block diagram illustration of the proposed melody extraction method.

The rest of the paper is organized as follows: The proposed melody extraction method is presented in Section 2. Evaluation and discussion of the results are presented in Section 3. Summary of the work and the possible future directions are presented in Section 4.

2. SPECTRO-TEMPORAL APPROACH TO MELODY EXTRACTION

The sequence of steps present in the proposed melody extraction method is illustrated in the form of a block diagram as shown in Fig. 1. The significance of each block is briefly explained in subsequent sections.

2.1. ZFF as a Bandpass Filter

A method to extract F0 from monaural speech signal by identifying epoch locations or instants of glottal closures (GCI) is presented in [11]. The method involves passing the speech signal through the cascade of two marginally stable, two pole, ideal digital filters resonating at Zero Hz (ZFR). Hence the resultant output is the polynomial function of time having exponential growing or decaying trend. In order to extract the epoch locations from the large values of fil-

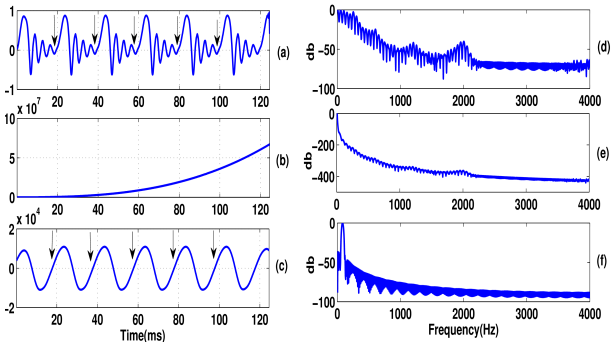

Fig. 2. Illustration of bandpass filtering nature of ZFF. The time domain waveforms of a segment of vowel, cascaded ZFR output, and the ZFF signal are shown in (a), (b) and (c) respectively. The corresponding spectrum of vowel, frequency-response magnitude of ZFR and ZFF signal are shown in (d), (e) and (f) respectively. The downward arrows in (a) and (c) represents the GCI.

tered output, each sample of the filtered output is subtracted by the running mean computed over a window of size equal to the average pitch period of the speech utterance considered. The instants of positive zero crossings of the mean subtracted signal (ZFF signal) is attributed to the locations of the impulse like excitation called epochs or GCI.

The time and frequency domain interpretations of the ZFF operations is presented in Fig. 2. A segment of synthetic vowel /a/, the output of the cascaded resonators and the ZFF signal are shown in Fig. 2(a), (b) and (c) respectively. The spectrum of vowel, cascaded ZFR log-magnitude response and spectrum of ZFF signal are shown in Figs. 2(d), (e) and (f) respectively. From the log-magnitude frequency-response of ZFR in Fig. 2(e), we can observe that the ZFR has mostly de-emphasised spectral information related to vocal tract and leaving behind a very significant emphasis at the Zero Hz. Also, from the spectrum of ZFF signal in Fig. 2(f) we can observe a strong peak around the region of pitch frequency. This effect can be attributed to the narrow bandpass (resonator like) filtering nature of mean subtraction window on the input speech signal.

2.1.1. Limitation of ZFF on music signals

In the original ZFF method, the signal used for extracting the F0 is the monaural speech signal consisting of single excitation source. Hence, time domain autocorrelation function is used to obtain the average pitch period to design the mean subtraction filter. On the other hand, music signal is a composite signal consisting of many pitched sources. Finding the resonant frequency or the average pitch period of the singer cannot be accomplished by using autocorrelation function. Also, ZFR is a marginally stable filter with two poles on the unit circle. Hence ZFF cannot be applied for the entire music signal considered because of the overflow due to finite precision of representation when representing exponential trend of the ZFR output. Furthermore, in the rendition of music signal, the pitch of the singer changes significantly from one note to the other note. Hence, it is not possible to use a single mean subtraction filter to derive the melody of the entire music signal. To support the above discussion, an excerpt of polyphonic music signal with note sequences which are rendered in ascending order and covering an octave is shown in Fig. 3. The waveform of the music excerpt is shown in Fig. 3(a), the corresponding spectrogram and the overlaid melody ground truth (blue contour) is shown is Fig. 3(b). The melody contours obtained by the original ZFF method (blue contours) with the mean subtraction window length set to the average pitch period calculated from

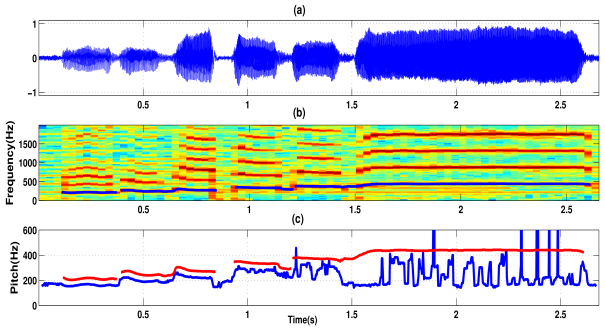

Fig. 3. Illustration of the limitation of the original ZFF on the music signal. (a) Music excerpt of note sequences covering an octave from 204Hz to 430Hz. (b) Spectrogram and the overlaid ground truth melody (blue contour). (c) Melody line obtained by original ZFF (blue contor) and the ground truth (red contour).

the autocorrelation function (which is about 5.7ms) and the ground truth (red contours) is shown in Fig. 3(c). The melody contour obtained by original ZFF method is shifted below 50Hz with respect to the ground truth for illustrative purpose. From Fig. 3(c) we can observe that the melody obtained from ZFF method from around 0.2s to 0.85s (i.e. first three notes) is intact with the ground truth. From fourth note onwards (around 0.95s onwards) we can observe that the ZFF is tracking spurious melody by tracking higher and lower octaves. Also from Fig. 3(c) we can observe that ZFF has accurately extracted the first three notes even though the resonance frequency which is supplied to the mean subtraction filter is equal to the frequency of neither of the notes. Hence, this property of ZFF which extracts the accurate melody even though the supplied resonance frequency is significantly away from the actual frequency of the melody, we call it as invariance property.

2.1.2. Invariance property of ZFF

In the previous subsection, we have introduced the invariability property of the ZFF. Here, we discuss the range of frequencies for which the ZFF is invariance with the help of spectrogram and the obtained melody contours. An excerpt of polyphonic music signal with complex melody modulation (from 0.4s-1.01s) is shown in Fig. 4. The waveform of the excerpt and the corresponding spectrogram with overlaid melody ground truth (blue contour) is shown in Fig. 4(a) and (b) respectively. ZFF is applied for the considered music segment (which has 5.8ms average pitch period) with the mean subtraction filters designed between 4ms to 7ms pitch period in-steps of 0.1ms. The corresponding melody contours for the selected pitch periods are shown in Fig. 5. Form Figs. 5(b), (c), (d) and (e), we can observe that the extracted melody contours exactly follows the ground truth starting from pitch period 4.5ms-6.5ms which spans approximately two pitch periods. In other words, we can accurately extract the melody from ZFF even if the supplied resonance pitch period (or frequency) of the mean subtraction filter is significantly away from the melody of the signal i.e., we have a greater flexibility of choosing the resonance frequency of the narrow bandpass filter for obtaining accurate melody.

2.2. Detection of Voiced and Unvoiced Segments

In Subsection 2.1 it is shown that the vocal source signal which is the impulsive excitation to the vocal tract system is emphasized in magnitude by passing the signal twice through the ZFR resonating at Zero Hz. Passing the signal twice through the ZFR has mostly attenuated the vocal tract resonances and hence significantly emphasized

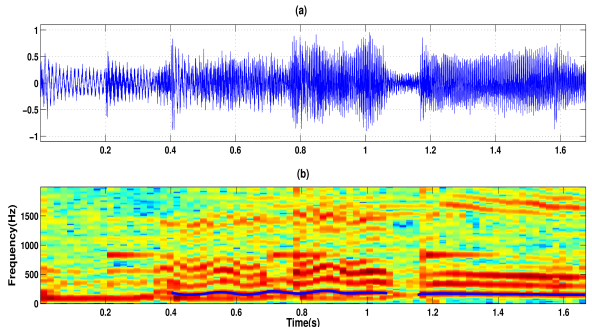

Fig. 4. An excerpt of a polyphonic music signal considered to show the invariance of the ZFF for a band of selected resonance frequencies. (a) Waveform of the music excerpt having complex melody modulation (0.4s-1.01s), (b) Spectrogram and the melody ground truth (blue)

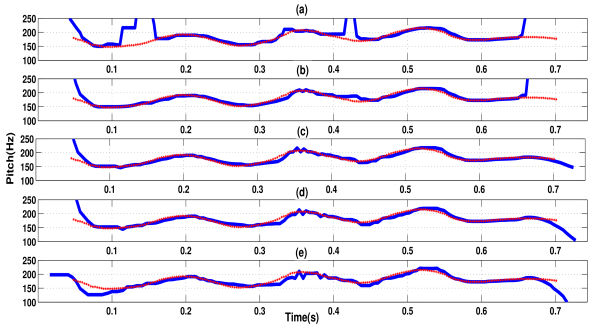

Fig. 5. Illustration of the invariance property of the ZFF for music excerpt considered in Fig.4 having a mean pitch period of 5.8ms in the singer modulation region from 0.4s to 1.01s. The output of the ZFF for various centre of frequencies shown in for (a) 4ms, (b) 4.5ms, (c) 5.4ms, (d) 5.9ms and (e) 6.5ms respectively.

the source signal. In a polyphonic music signals with lead voice, it is the vocal source which is mostly dominant when compare to the other sources. Hence by exploiting the strength of excitation (SOE) of the source signals, the vocalic regions can be identified. Initially, the composite signal is zero frequency filtered with a window size of 1ms (1000Hz). The window size of 1ms is chosen since it covers the entire frequency range of the vocal source. The ZFF allows a composite source signal which is the sum of the sources i.e., vocal source and the sources of the other instruments. Because of the dominance property of the vocals, the ZFF signal has high energy in the voiced regions and very low energy in the unvoiced regions. The strength of excitation contour is obtained as the slope of the ZFF signal at the instants of zero crossings of the ZFF signal. SoE is passed through the Savitzky-Golay filter of order 3 and window size of 31 samples to obtain the smoothed envelope.

An excerpt of a polyphonic music signal, its spectrogram, spectrogram of the ZFF signal and smoothed SoE with overlaid detected voiced boundary markers is shown in Fig. 6. From Fig. 6(c) we can observe the increased energy of the ZFF signal of Fig. 6(a) in the harmonics of the source signal when compared to the original spectrogram of Fig. 6(b). Though the energy in other parts of the spectrogram is also increased but it is not predominant as compared to the energy of the harmonic of the source signal. Furthermore, from Fig. 6(d), we can observe a large threshold range available for voiced and unvoiced (VUV) decision. The mean μ_{SoE} and standard deviation σ_{SoE} of the smoothed SoE is computed. A threshold based on statistical measure μ_{SoE} - $\delta * \sigma_{SoE}$ is computed for VUV classification. Where δ is the deviation parameter, an optimum value of 0.95 is chosen to reduce false alarms.

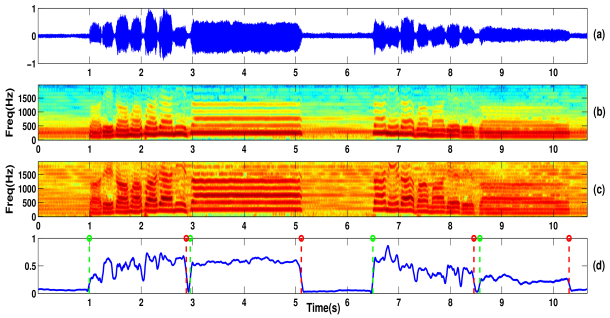

Fig. 6. Illustration of the effect of ZFR on composite music signal. (a) Music excerpt, (b) the Corresponding spectrogram, (c) Spectrogram of ZFF signal and (d) SoE contour and voiced segment markers (vertical markers).

2.3. Detection of Voiced Note Onsets

The fundamental frequency of the melody source varies significantly from one note to other. Hence, a single mean subtraction filter is not sufficient to remove the trend in the ZFR output to obtain the accurate F0 of the lead voice from the entire music signal. To overcome this limitation, the voiced segments identified in the previous subsection are further segmented into voiced note like regions by identifying note onsets. An onset can be defined as an event in a music signal where the signal properties such as short time energy, spectral magnitude, phase spectrum etc., shows significant changes [12, 13, 14, 15, 16]. Music signal with lead voice consists of both hard and soft onsets. The hard onsets are characterised by high frequency energy, therefore the linear frequency weighted energy content analysis for the sub-band spanning from 1KHz-10KHz is applied. Then the onset detection function is obtained by the derivative of the energy function which showed sharp peaks at the instants of note onsets.

On the other hand, the soft onsets are characterised by changes in the frequency content specifically at lower frequency band spanning 50Hz-1.5KHz. Hence, a method similar to [17] is followed to determine the soft onsets. To determine the spectral changes, Euclidean distance between the spectral frames is measured as

$$E_{dm}(n) = \sum_{k;E_X(n,k)>0} E_X(n,k)^2 \quad (1)$$

where

$$E_X(n,k) = X(k,n) - X(k,n-1) \quad (2)$$

The distance measure is normalized to detect soft onsets along with hard onsets given by

$$E_{dm(norm)}(n) = \frac{E_{dm}(n)}{\sum_{k=f1}^{f2} |X(k,(n-1))|^2} \quad (3)$$

To suppress the noisy regions in the detection functions which leads to multiple onset detection without blurring the position of onsets and smoothing weaker onsets. In time domain, a low pass filtering is performed by taking the difference between the current frame and the contribution of exponentially weighted previous frames of detection function, given by

$$y(n) = F(n) - \sum_{a=1}^{A} \frac{F(n-a)}{a} \quad (4)$$

where $F(n)$ represents the onset detection functions determined previously for soft and hard onsets and a is the weighting factor. The

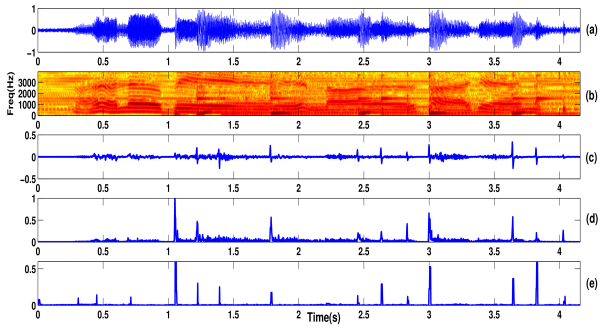

Fig. 7. Illustration of the onset detection functions of a polyphonic music signal. (a) Waveform (b) Spectrogram, (c) and (d) Hard and Soft onset detection functions and (e) Combined and smoothed onset detection functions of (c) and (d).

filtered onset detection functions are combined and the location of onsets are obtained by peak picking heuristics as follows: The n^{th} frame is considered as onset if the onset detection function fulfils the following conditions

$$y(n) = max(y(n-w)) \tag{5}$$

$$y(n) >= mean(y(n-w:n+w)) + \delta \tag{6}$$

$$n - n_{lastonset} > w \tag{7}$$

The optimal values for parameters w, δ and A are chosen as 3, 0.05 and 20 respectively. The hard, soft and the combined smoothed onset detection functions of an excerpt of polyphonic music signal are shown in Fig. 7. From Fig. 7(c) we can observe that the spikes in the detection function corresponds to the vertical ridges in the spectrogram which are hard onsets resultant of the wideband spectral energy. The spikes in the Fig. 7(d) which corresponds to both hard and soft onsets which are due to the normalized low frequency spectral changes. The smoothed combined detection functions in Fig. 7(e) shows almost zero noise variance which in-turn helps in picking the low energy peaks of soft onsets.

2.4. Resonance Frequency Detection and Adaptive Filtering

To obtain the melody of lead voice in the polyphonic music signal, the trend in the output of the ZFR of each note segment should be removed adaptively with the mean subtraction window length corresponding to the average pitch period of the lead voice in the segment considered. Since the polyphonic music signal consists of several pitched instruments, the average pitch period of voice source in a note segment is obtained by TWM error function [18]. TWM error function is designed to find the F0 of the given signal by minimizing the error between the measured partial peaks and the predicted harmonics. The measured partial peaks are obtained from short time Fourier transform (STFT) with 40ms frame size and 3ms frame shift by sinusoidal detection proposed in [19]. The sinusoids in a STFT frame are determined by measuring the mean squared error difference between measured spectral peak's shape and the spectrum of the analysis window main lobe. Then the probable (predicted) F0 candidates for TWM are obtained as the sub-multiples of measured sinusoids. The F0 search range is limited to 50Hz-1KHz assuming that voice F0 will remains in this range. The representative pitch period of a note segment is obtained as the reciprocal of the average of F0 candidates for which the TWM error is minimum.

Finally, Zero frequency filtering is performed on each identified note segments separately with the trend removal window designed with the average pitch period of the corresponding note segment. The instants of zero crossings of all note segments which represents the GCI's are obtained and the inverse of the difference between successive GCI's are computed to obtain the melody of lead voice.

3. EVALUATION AND DISCUSSION

The proposed melody extraction method is evaluated using three openly available datasets which includes music excerpts and the corresponding F0 ground truth in the form of time-frequency pairs: ADC2004, Mirex05TrainFiles and MIR-1K dataset are considered for evaluation. Which consists of 20, 13 and a subset of 280 excerpts and each having duration between 7-40s in the genres of pop, jazz, opera, rock, solo classical piano sung by both male and female singers. The four global measures provided in MIREX 2005 [20] are used for evaluating the proposed method: **Voicing Recall Rate** (VR), **Voicing False Alarm Rate** (VFA), **Raw Pitch Accuracy** (RP) and **Overall Accuracy** (OA). The performance of the proposed method is compared with widely used and openly available salience based melody extraction system Melodia[1] [10] as shown in Table 1. From Table 1 we can observe that the performance of the proposed method is comparable with that of Melodia. A slight decrease in the VFA of the proposed method is observed. It can be attributed to the large dynamic range of the SoE contour for threshold even in the regions of the music signal where the strengths of accompaniment and the vocals are comparable. Hence the increase in VR, RP and OA performance for both ADC2004 and Mirex05TrainSet. A decrease in overall accuracy of the proposed method is observed in the MIR-1K dataset this is due to the tracking of the higher octaves in the regions of strongly excited percussive regions. Presently, the performance of the proposed method is comparable to that of the Melodia, but the performance can be significantly improved by preprocessing the music signal to suppress the percussion component. Due to the time-frequency uncertainty of the STFT, the note onsets and offsets are either detected later or earlier then the true locations which in turn contributed to the performance accuracy. Also, the performance of the proposed method is significantly affected by the resonance frequency of the mean subtraction filter which is sometimes detected beyond the invariance range by TWM algorithm. More results of extracted melody (including the failure cases) on various polyphonic music signals are provided at `https://github.com/mgurunathreddy/Melody-extraction-results`.

Table 1. Performance comparison of proposed (P) method and Melodia (M).

Dataset	VR		VFA		RP		OA	
	P	M	P	M	P	M	P	M
ADC2004	0.81	0.79	0.18	0.21	0.78	0.75	0.74	0.72
Mirex05TrainSet	0.80	0.77	0.19	0.23	0.73	0.69	0.70	0.67
MIR-1K	0.79	0.89	0.22	0.19	0.77	0.84	0.76	0.82

4. SUMMARY AND CONCLUSIONS

A mixed time and frequency domain melody extraction method is proposed by exploiting the bandpass filtering nature of the ZFF. The voiced and unvoiced regions of the polyphonic music signal are detected by thresholding SoE contour. The voiced note like segments are obtained by detecting note onsets. Finally, the melody contour is extracted by filtering each voiced note segment with adaptive zero frequency filtering. The initial results obtained from the proposed method are comparable to that of the state of the art melody extraction system which is really encouraging to improve the accuracy of the proposed method in near future. Also we would like to compare the proposed method with other salience and source separation methods with larger datasets.

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

5. REFERENCES

[1] J. Salamon, E. Gomez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[2] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[3] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 425–428.

[4] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 57–60.

[5] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 73–84, 2013.

[6] M. P. Ryynänen and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[7] M. 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.

[8] R. P. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness," *Computer Music Journal*, vol. 30, no. 4, pp. 80–98, 2006.

[9] 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.

[10] 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.

[11] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

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

[13] S. Dixon, "Onset detection revisited," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2006, pp. 133–137.

[14] P. Leveau and L. Daudet, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Proceedings of the International Symposia on Music Information Retrieval (ISMIR)*, 2004.

[15] B. Scherrer and P. Depalle, "Onset time estimation for the analysis of percussive sounds using exponentially damped sinusoids," in *Proceedings of the Internatonal Conference on Digital Audio Effects (DAFx)*, 2014, pp. 211–217.

[16] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods." in *Proceedings of International Symposia on Music Information Retrieval (ISMIR)*, 2012, pp. 49–54.

[17] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, 2002, pp. 33–38.

[18] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[19] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[20] G. E. Poliner, D. P. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.