INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Enhanced Harmonic Content and Vocal Note Based Predominant Melody Extraction from Vocal Polyphonic Music Signals

Gurunath Reddy M, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology, Kharagpur, India
{mgurunathreddy, ksrao}@sit.iitkgp.ernet.in

Abstract

A method based on the production mechanism of the vocals in the composite vocal polyphonic music signal is proposed for vocal melody extraction. In the proposed method, initially the non-pitched percussive source is suppressed by observing its wideband spectral characteristics to emphasise the harmonic content in the mixture signal. Further, the harmonic enhanced signal is segmented into vocal and non-vocal regions by thresholding the salience energy contour. The vocal regions are further divided into vocal note like regions by their spectral transition cues in the frequency domain. The melody contour in each vocal note is extracted by detecting the locations of instant of significant excitation by passing it through adaptive zero frequency filtering (ZFF) in the time domain. The experimental results showed that the proposed method is indeed comparable to the state-of-the-art saliency based melody extraction method.
Index Terms: Predominant Melody, Zero Frequency Filter, Note Onsets, Vocal Notes, Polyphonic Music, Vocals and Non-Vocals.

1. Introduction

Predominant melody is the single fundamental frequency (F0) contour of the dominant instrument in the polyphonic music signal [1]. The dominant instrument can be either a human singing voice or any lead instrument. Since the majority of the available polyphonic music signals contain vocals as dominant source, vocal melody extraction is the goal of this paper. The extracted melody can be used in many potential applications [2], such as query by humming [3], singer identification [4], automatic music transcription [5], music genre classification [6], dominant instrument identification, cover song detection, music desoloing [7] and so on.

We can broadly classify the available melody extraction methods into two categories viz. (1) Signal transformation (salience) and (2) Source separation based methods. Signal transformation is a separation less method in which mostly polyphonic music signal is transformed into spectral domain by short-time-Fourier-transform (STFT). Followed by estimating the pitch saliency function by summation of harmonic partials. Finally, melody contour tracking algorithms are applied on the candidate pitches obtained from the salience function. Salience based methods mostly differ in the following aspects : pitch saliency function computation, salience peak estimation and melody contour creation from the candidate pitches [8, 9, 10, 11, 12]. On the other hand, source responsible for melody in the polyphonic music signal is separated from the rest of mixture signal in separation based methods. Melody is extracted from the separated source signal by a monophonic pitch detection algorithm [13, 14, 15, 16]. A detailed review on available salience, source separation and other melody extraction methods can be found in [2].

In this paper, the digital source-filtering model [17] of the speech production mechanism is adopted for the melody extraction from the music signal. Though the speech and vocal polyphonic music are entirely different signals, still they share the common production mechanism. That is, the major source of excitation is the impulse-like excitation to the time varying vocal-tract system in speech and vocals in vocal polyphonic music signals. The impulse excitation to the system results in discontinuity in the frequency of the output signal produced. The discontinuity due to impulse is reflected across all frequencies including the zero frequency. Further, the frequency near zero Hz should essentially contains the information about the impulsive excitation. Hence, in this work, the zero frequency filtering [18] method is adopted for extracting the instants of significant excitation or glottal closure instants (GCIs) from the vocal music signal. Originally, ZFF is proposed to extract the GCIs by passing the monophonic speech signal through a cascade of two zero frequency resonators (ZFRs). Followed by designing a mean subtraction filter whose length in samples equal to the average pitch period estimated from the autocorrelation function to extract the GCIs by removing the trend in the output of the ZFRs. Further, the instantaneous F0 is computed as the reciprocal of the distance between the consecutive GCIs. The ZFF method as it is cannot be applied for the polyphonic music signals because of the following reasons: (i) it consists of many pitch and non-pitched sources, (ii) the melody of the singer varies significantly from one note to the other, (iii) the source of excitation of non-pitched percussive instrument is impulsive like, (iv) unlike speech, the coupling of the source and the filter in vocals is very strong. Hence, in this method, initially the percussive component in the polyphonic signal is suppressed by observing the wideband spectral characteristics in the frequency domain. The percussion suppressed signal is segmented into vocal and non-vocal regions by thresholding the harmonic partials energy contour. Further, the vocal regions are divided into vocal note like regions by finding their onsets in the frequency domain. Finally, each note is adaptively zero frequency filtered after suppressing the strong source-system coupling and designing a narrow bandpass filter with resonance frequency obtained from the Two-way-miss-match (TWM) algorithm to construct the melody contour.

2. Source-Filter Model Based Melody Extraction Method

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.

http://dx.doi.org/10.21437/Interspeech.2016-856

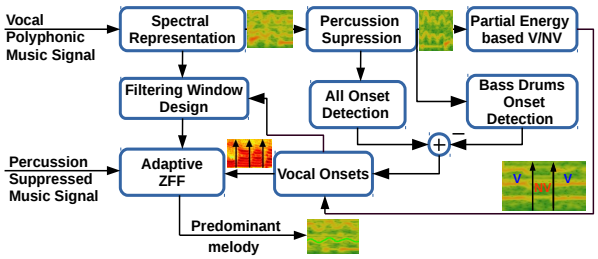

Figure 1: *Block diagram illustration of the proposed melody extraction method.*

2.1. ZFF as a Source-Filter Separator

A method to extract F0 from monaural speech by separating the signal containing the excitation source and the filter information is proposed in [19]. The basis of the proposed method is, the discontinuity due to impulse like excitation effects all frequencies equally, including the frequencies near zero Hz. Hence, the output of the ZFR essentially contains the information about the discontinuities due to impulse-like excitation. In order to separate the signal containing the excitation information, (i) the speech signal is passed twice through the ZFR given by $y_0[n] = \sum_{k=1}^{4} a_k y_0[n-k] + x[n]$, where $a1 = 4$, $a2 = -6$, $a3 = 4$, and $a4 = -1$, (ii) to find the overriding epoch locations, the trend in each sample of the signal $y_0[n]$ is removed by subtracting the mean computed over a window length equal to the average pitch period of the speaker given by $y[n] = y_0[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_0[n+m]$, (iii) the GCIs are obtained as the positive zero crossings of the ZFF signal ($y[n]$), and (iv) the instantaneous pitch contour is computed as the reciprocal of the difference between successive GCIs.

Time and frequency domain interpretations of the ZFF is illustrated in Fig. 2. A segment of synthetic vowel /a/, the output of the cascaded ZFR and the ZFF signal are shown in Fig. 2(a), (b) and (c) respectively. The corresponding spectrum of vowel, magnitude response of cascaded ZFR and ZFF signal are shown in Fig. 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-emphasises spectral information related to vocal tract and very significant emphasis near the zero Hz in terms of magnitude. 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 filter (MSF) on the ZFR output containing the overriding information about GCIs. The mean subtracted signal in Fig. 2(c) is essentially a single low frequency signal, whose positive zero crossings corresponds to the instants of glottal closures. The GCI locations do not deviate significantly as long as the obtained average pitch period is within 1-2 pitch period of the speaker for MSF, which we call it as invariance property of ZFF.

2.2. Percussion Suppression for Harmonic Enhancement

The harmonic content in the vocal polyphonic music signal is enhanced by suppressing the wideband spectral energy of the non-pitched percussive instrument (NPPI). The NPPI not only interfere with the harmonic partials of the pitched instruments, but also frequency content near zero Hz. Hence, the wideband spectral energy is suppressed by computing the frequency change in the STFT of the polyphonic music signal. The polyphonic music signal is transformed to frequency domain by STFT of 40ms frame size and 3ms frame shift. A relatively small frame shift of 3ms is chosen to retain the time resolution of rapidly decaying percussive source along the time. For each

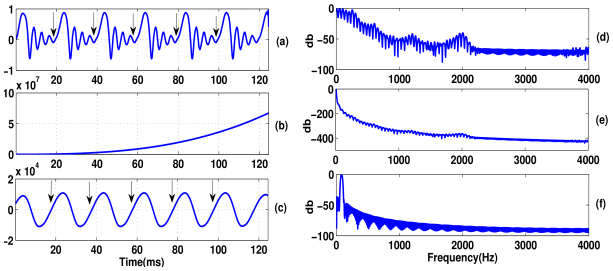

Figure 2: *Illustration of ZFF as a source-filter separator. The time domain waveforms of a segment of vowel, cascaded ZFR output, and the ZFF signal are shown in (a), (b) and (c) respectively. The corresponding spectrum of vowel, magnitude response of cascaded ZFR and ZFF signal are shown in (d), (e) and (f) respectively. The GCIs are shown as downward arrows in (a) and (c).*

signal frame, the STFT is computed as

$$F(l,k) = \sum_{n=0}^{N-1} x(n)w(n)e^{-j2\pi kn/N} \quad (1)$$

where $F(l,k)$ is the l^{th} frame, k^{th} frequency complex spectral frame, $x(n)$ is the music signal, $w(n)$ is the hamming window, N = 2048 is the number of frequency bins. The wideband spectral energy is suppressed by taking the frequency change in the magnitude spectrum of $F(l,k)$ by

$$X_{fc}(l,k) = X(l,k) - X(l,k-1) \quad (2)$$

where $X(l,k)$ is the magnitude spectrum of $F(l,k)$. The harmonic content of the spectrum is retained and enhanced by

$$X_{pow}(l,k) = X_{fc}(l,k)^2|_{X_{fc}(l,k)>0} \quad (3)$$

A binary mask is created to suppress the percussion from each spectral frame by

$$X_{mask}(l,k) = X_{pow}(l,k) > (\underset{(l,k)}{\text{argmax}}(X_{pow}(l,k)) * \delta/100 \quad (4)$$

where δ is the parameter decides the amount of harmonic partials needs to be retained. An optimal value of 0.1 is chosen for δ to retain the maximum amount of harmonic partials. The magnitude spectrum $X_{pow}(l,k)$ is smoothed with a five point median filter to remove any isolated peaks in spectrum given by

$$X_{med}(l,k) = medfilt(X_{pow}(n-l:n+l,k)) \quad (5)$$

The binary mask $X_{mask}(l,k)$ is multiplied with the magnitude $X_{med}(l,k)$ and phase $P(l,k)$ (phase of Eq. 1) spectrum to get the percussion suppressed magnitude and phase spectrum given by

$$X_{mod}(l,k) = X_{med}(l,k) * X_{mask}(l,k) \quad (6)$$

$$P_{mod}(l,k) = P(l,k) * X_{mask}(l,k) \quad (7)$$

The harmonic enhanced polyphonic signal is obtained by inverse STFT given by

$$y[n] = 1/N \sum_{n=0}^{N-1} X_{mod}(l,k)e^{-jP_{mod}(l,k)}e^{j2\pi kn/N} \quad (8)$$

An illustration of percussion suppression is shown in Fig. 3. Fig. 3(a) is the spectrogram of the polyphonic music signal consists of both harmonic and wideband percussion source (shown in ellipses). Figs. 3(b) and (c) shows the percussion suppressed and median filtered spectrograms. From Fig. 3(c) we can observe that the wideband percussion is mostly suppressed and harmonic component in the spectrogram is significantly enhanced.

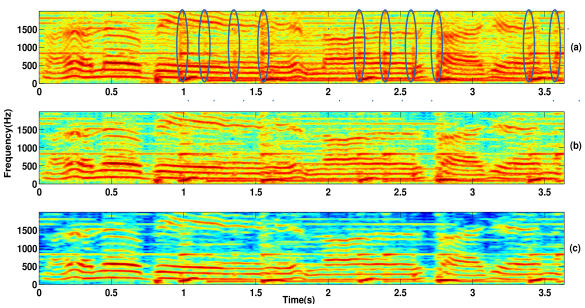

Figure 3: *Illustration of the percussion suppression of polyphonic music signal. (a) Polyphonic music signal containing percussive and harmonic sources, (b) percussion suppressed magnitude spectrogram, (c) median filtered and hence harmonic enhanced spectrogram.*

2.3. Vocal/Non-Vocal Detection

The vocal and non-vocal (V/NV) refers to the vocal melody and non-melody regions in the polyphonic music signal. The dominant harmonic partials in the percussion suppressed median filtered magnitude spectrum $X_{mod}(l,k)$ in the frequency range 100 Hz - 4 KHz (vocal activity ceases above 4 KHz) is obtained by comparing with the maximum partial peak. The partials with less than $1/10$ of the maximum peak is filtered out from each frame. For all frames, the mean μ_{dp} and standard deviation σ_{dp} of all dominant partial is computed. The partials with magnitude below μ_{dp} - δ_{dp} σ_{dp} are removed from all frames in order to give emphasis to the dominant partials. The energy of the remaining dominant partials in each frame is computed as

$$E[l] = \sum_{k}^{P} X_{rem}(l,k)^2 \tag{9}$$

Where $X_{rem}(l,k)$ contains the remaining dominant harmonic partials. The energy contour $E[l]$ is passed through the Savitzky-Golay filter [20] of order 3 and window size 31 frames to obtain the smoothed envelope. An excerpt of polyphonic music signal, smoothed energy contour with overlaid detected vocal boundary markers is shown in Fig. 4. The mean μ_E and standard deviation σ_E of the smoothed energy contour is computed. The regions of energy contour is labelled as vocal for which energy is greater than the statistical measure μ_E - $\delta_E * \sigma_E$. Where δ_E is the threshold deviation parameter, an optimum value of 0.95 is chosen to reduce miss rates.

2.4. Vocal Note Onset Detection

The vocal melody varies significantly from one note to the other. Hence, a single MSF is not sufficient to remove the trend in ZFR output of the entire music signal. Therefore, the vocal regions are further divided into vocal note like regions by detecting their onsets in the median filtered magnitude spectrogram. 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 [21, 22, 23, 24, 25]. The vocal onsets are manifested as both soft and hard onsets in the lower frequency range. Hence, the onsets are detected as spectral changes in the vocal frequency range spanning 100 Hz-4 KHz. A method similar to [26] is adopted to determine the spectral changes by finding the Euclidean distance between the spectral frames given by

$$E_{dm}(l) = \sum_{k;E_x(l,k)>0} E_x(l,k)^2 \tag{10}$$

where

$$E_x(l,k) = X_{mod}(l,k) - X_{mod}(l-1,k) \tag{11}$$

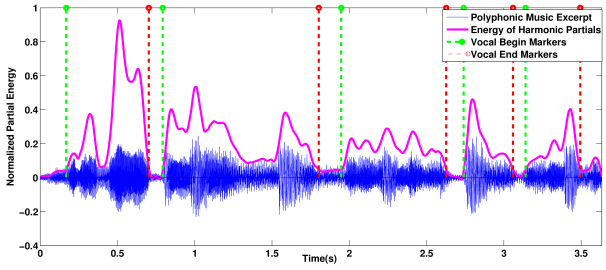

Figure 4: *Illustration of the polyphonic music signal with the smoothed harmonic partial energy contour and overlaid vocal segment boundary markers.*

The distance measure is normalized to obtain the onset detection function whose peaks correspond to the onsets given by

$$E_{dmn}(l) = \frac{E_{dm}(l)}{\sum_{k=f1}^{f2} X_{mod}(l-1,k)^2} \tag{12}$$

The onset detection function $E_{dmn}(l)$ contains peaks corresponding to vocal notes as well other pitched percussive instruments (bass and snare onsets). Resulting in segmenting a note into several sound units. In order to suppress the other peaks, the spectral change along time in the same vocal frequency range is computed on the frequency differenced spectrogram as follows. The median filtered spectrogram $X_{mod}(l,k)$ is exponentially weighted to emphasize the low frequency onsets such as the bass and snare.

$$X_w(l,k) = \sum_{k=f1}^{f2} 1/k * X_{mod}(l,k) \tag{13}$$

The weighted frequency difference is taken along the frequency axis given by

$$X_{fd}(l,k) = X_w(l,k) - X_w(l,k-1) \tag{14}$$

The spectral change of $X_{fd}(l,k)$ along the time is taken to remove the harmonics and hence to retain the pitched percussive onsets given by

$$X_{sc}(l,k) = X_{fd}(l,k) - X_{fd}(l-1,k) \tag{15}$$

The normalized energy of the positive spectral changes for each frame along the time is computed to obtain the onset detection function given by

$$X_{df}(l) = \frac{\sum_{k;E_{fd}(l,k)>0} X_{fd}(l,k)^2}{\sum_{k=f1}^{f2} X_w(l-1,k)^2} \tag{16}$$

The location of onsets in the onset detection function of Eqs. 12 and 16 are obtained by peak picking heuristics as follows: The l^{th} frame is considered as onset if the onset detection function fulfils the following conditions (here, $y(l)$ can be either $E_{dmn}(l)$ or $X_{df}(l)$)

$$y(l) = max(y(l-w)) \tag{17}$$

$$y(l) >= mean(y(l-w:l+w)) + \delta \tag{18}$$

$$l - l_{lastonset} > w \tag{19}$$

The optimal values for w and δ are chosen as 3 and 0.05 respectively. The location of final onsets detected from 16 are removed from the set containing the onset locations of Eqs. 12 which are at a distance of four frames to mostly retain the vocal onsets. The process of vocal note onset detection is illustrated in Fig. 5. Fig. 5(a) shows the spectrogram containing the pitched percussive note onsets and its onset detection function in Fig. 5(b). The spectral change based onset detection function and the final vocal note onsets are shown in Fig. 5(c) and (d) respectively.

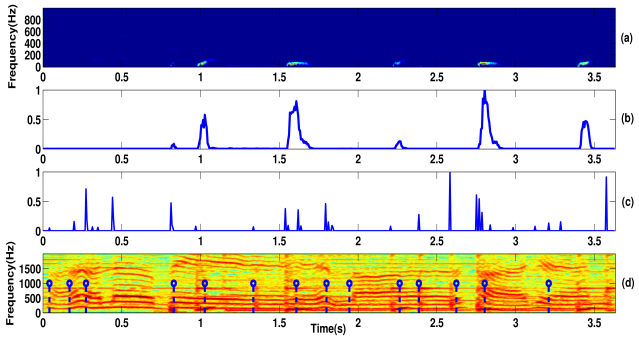

Figure 5: *Illustration of onset detection functions of a music excerpt. (a) Spectrogram showing the signature of pitched percussive regions obtained from frequency differenced spectrogram (b) onset detection function of (a), (c) spectral change based onset detection function, and (d) spectrogram and the overlaid final vocal onsets.*

2.5. Resonance Frequency Detection and Adaptive Filtering

The melody contour in each vocal note is obtained by extracting the GCIs by adaptive zero frequency filtering. In order to remove the trend in the output of the ZFR of each vocal note, an average pitch period or center of frequency of the respective vocal note for designing the narrow bandpass filter or MSF is obtained by TWM algorithm [27]. 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 in each STFT frame. For each frame, the measured partial peaks are obtained from the percussion suppressed and median filtered spectrogram $X_{mod}(l,k)$ by sinusoidal detection [28]. The sinusoids in each frame is obtained by measuring a mean squared error difference between measured spectral peak's shape and the spectrum of the analysis window main lobe. The probable (predicted) F0 candidates for TWM algorithm are obtained as the sub-multiples of measured sinusoids. The F0 search range is limited to 50 Hz-1 KHz assuming that the vocal melody will lie in this range. The representative pitch period of vocal note is obtained as the reciprocal of the median of F0 candidates for which the TWM error is minimum.

In order to strongly de-emphasize the system resonances due to vocal tract and instruments, and hence to emphasize the source information. Each vocal note of the percussion suppressed polyphonic music signal $y[n]$ of Eq. 8 is passed through the cascade of three ZFRs given by

$$Y[n] = \sum_{k=1}^{6} a_k Y[n-k] + y[n] \qquad (20)$$

where, $a1 = 6, a2 = -15, a3 = 20, a4 = -15, a5 = 6, and a6 = -1$. The trend in the cascaded ZFR output is removed by filtering the signal twice through the mean subtraction filter (as discussed in subsection 2.1) designed with the center of frequency computed from TWM algorithm for respective vocal note. Finally, the GCIs of the trend removed signal i.e., the ZFF signal is obtained as the negative to positive zero crossings. The melody is computed as the reciprocal of the difference between successive GCIs.

3. Evaluation and Discussion

The performance of the proposed melody extraction method is evaluated on three openly available datasets. The datasets includes music excerpts and the corresponding melody ground truth in the form of time-frequency pairs. ADC2004, Mirex05TrainFiles and MIR-1K datasets are considered for evaluation, consisting of 20, 13 and a subset of 400 excerpts respectively. Each excerpt had a duration between 7 - 40 sec in the genres of pop, jazz, opera, rock, solo classical piano sung by both male and female singers. The four global measures provided by MIREX 2005 [1] 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 saliency based melody extraction method Melodia[1] [12] as shown in Table 1. From Table 1 we can observe that the performance of the proposed method is indeed comparable with that of Melodia for the dataset considered.

The overall increase in the performance of the proposed method is observed for the datasets ADC2004 and Mirex05TrainSet. The increase in performance is mainly due to the percussion suppression resulted in harmonic rich music excerpts benefited by the TWM algorithm for identifying the resonance frequency within the invariance range at each vocal note. And hence, ZFF in succeeded in extracting the correct GCIs. The overall increase in VFA is observed for all datasets this is mainly due to occasionally misclassification of vocals as non-vocals due to sensitivity of the threshold in the strong pitched percussive regions. Overall decrease in the performance of the proposed method is observed compared to the Melodia for a slightly larger dataset MIR-1K mostly due to the tracking of the representative resonance frequency by the TWM algorithm beyond the invariance range of MSF. In future, we would like to address the sensitivity of the threshold for V/NV classification by adaptive thresholding techniques. A modified TWM algorithm for extracting the resonance frequency within the invariance range by constraining error computation based on the dominant harmonic partials. Also, the proposed method needs to be evaluated on the larger dataset covering various genre and styles other than the considered dataset.

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

Dataset	VR		VFA		RP		OA	
	P	M	P	M	P	M	P	M
ADC2004	0.83	0.79	0.22	0.21	0.80	0.75	0.76	0.72
Mirex05TrainSet	0.82	0.77	0.25	0.23	0.75	0.69	0.74	0.67
MIR-1K	0.81	0.84	0.24	0.17	0.79	0.85	0.78	0.81

4. Summary and Conclusions

A predominant vocal melody extraction method based on GCIs of the vocal source signal is proposed. The influence of the non-pitched percussive source on the mixture signal is suppressed by its wideband spectral characteristics to emphasise the harmonic content in the polyphonic signal. The harmonic enhanced signal is further segmented into vocal and non-vocal regions by thresholding the partial harmonic energy contour. The vocal regions are further divided into vocal note like regions by their spectral transition cues in the frequency domain. The melody contour in each vocal note is extracted by detecting GCIs by passing it through an adaptive zero frequency filtering (ZFF) in time domain. The experimental results showed that the proposed method is indeed comparable to the state-of-the-art saliency based melody extraction method for the datasets considered.

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

5. References

[1] 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.

[2] 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.

[3] J. Salamon, J. Serra, and E. Gómez, "Tonal representations for music retrieval: from version identification to query-by-humming," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 1, pp. 45–58, 2013.

[4] R. Foucard, J.-L. Durrieu, M. Lagrange, and G. Richard, "Multimodal similarity between musical streams for cover version detection," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5514–5517.

[5] E. Gómez, F. J. Cañadas-Quesada, J. Salamon, J. Bonada, P. V. Candea, and P. C. Molero, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing." in *Proceedings of International Symposia on Music Information Retrieval (ISMIR)*, 2012, pp. 601–606.

[6] J. Salamon, B. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 81–84.

[7] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 105–108.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.

[18] B. Yegnanarayana and K. Sri Rama Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[19] 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.

[20] R. W. Schafer, "What is a savitzky-golay filter?[lecture notes]," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.