SINGING VOICE ANALYSIS AND EDITING BASED ON MUTUALLY DEPENDENT F0 ESTIMATION AND SOURCE SEPARATION

Yukara Ikemiya *Kazuyoshi Yoshii* *Katsutoshi Itoyama*

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

This paper presents a novel framework that improves both vocal fundamental frequency (F0) estimation and singing voice separation by making effective use of the mutual dependency of those two tasks. A typical approach to singing voice separation is to estimate the vocal F0 contour from a target music signal and then extract the singing voice by using a time-frequency mask that passes only the harmonic components of the vocal F0s and overtones. Vocal F0 estimation, on the contrary, is considered to become easier if only the singing voice can be extracted accurately from the target signal. Such *mutual* dependency has scarcely been focused on in most conventional studies. To overcome this limitation, our framework alternates those two tasks while using the results of each in the other. More specifically, we first extract the singing voice by using robust principal component analysis (RPCA). The F0 contour is then estimated from the separated singing voice by finding the optimal path over a F0-saliency spectrogram based on subharmonic summation (SHS). This enables us to improve singing voice separation by combining a time-frequency mask based on RPCA with a mask based on harmonic structures. Experimental results obtained when we used the proposed technique to directly edit vocal F0s in popular-music audio signals showed that it significantly improved both vocal F0 estimation and singing voice separation.

***Index Terms*—** Vocal F0 estimation, singing voice separation, melody extraction, robust principal component analysis (RPCA), subharmonic summation (SHS).

1. INTRODUCTION

Active music listening [1] has recently been considered one of the most attractive directions in music signal processing research. While listening to music, we often wish that a particular instrument part were performed in a different way. Such a music touch-up is generally infeasible for commercial CD recordings unless individual instrument tracks are available, but the state-of-the-art techniques of music signal processing enable us to *actively* make small changes to existing CD recordings with or without using score information. Drum parts, *e.g.*, can be edited in MIDI sequencers [2], and the volume balance between multiple instruments can be adjusted [3, 4].

Since the sung melody is an important factor affecting the mood of popular music, several methods have been proposed for analyzing and editing the three major kinds of acoustic characteristics of the singing voice: pitch, timbre, and volume. Ohishi *et al.* [5], for example, proposed a method that represents the temporal dynamics of a vocal F0 contour by using a probabilistic model and transfers those dynamics to another contour. A similar model was applied to a volume contour of the sung melody. Note that those methods can deal only with isolated singing voices. Fujihara and Goto [6], however, proposed a method that can be used to directly modify the spectral envelopes (timbres) of the sung melody in a polyphonic music audio signal without affecting accompanying instrument parts.

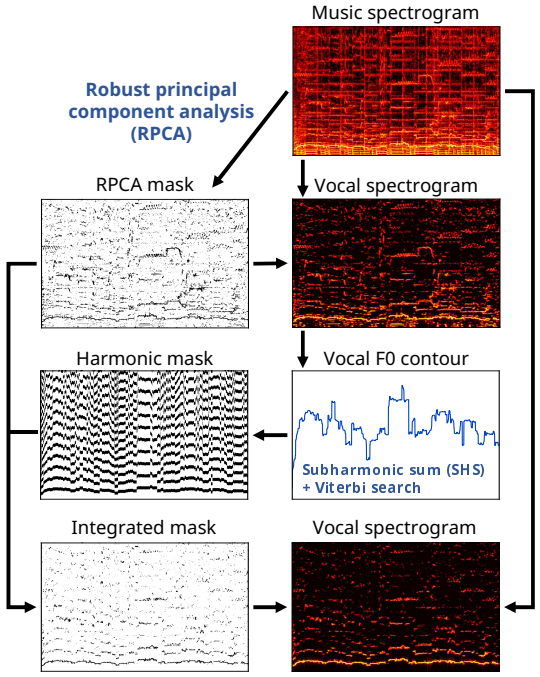

Fig. 1. Overview of proposed framework

To develop a system that enables users to edit the acoustic characteristics of the sung melody included in a polyphonic mixture, we need to perform accurate vocal F0 estimation and singing voice separation. Although these two tasks are intrinsically linked with each other, only the *one-way* dependency between them has conventionally been considered. A typical approach to vocal F0 estimation is to identify a series of predominant harmonic structures from a music spectrogram [7–9]. Salamon and Gómez [10] focused on the characteristics of vocal F0 contours to distinguish which contours derived from vocal sounds. To improve vocal F0 estimation, some studies used singing voice separation techniques [11–13]. This approach is effective especially when the volume of the sung melody is relatively low [14]. A typical approach to singing voice separation is to use a time-frequency mask that passes only the harmonic components of vocal F0s and overtones [15–17]. Several methods do not use vocal F0 information but instead, focus on the repeating nature of accompanying sounds [13, 18] or the spectral characteristics of the sung melody [11, 19]. Durrieu et al. [20] used source-filter NMF for directly modeling the F0s and timbres of singing voices and accompaniment sounds and separating each type of sounds.

This study was partially supported by JSPS KAKENHI 26700020, 24220006, 24700168 and CREST OngaCREST project.

In this paper we propose a novel framework that improves both vocal F0 estimation and singing voice separation by making effective use of the *mutual* dependency of those two tasks. The proposed method of singing voice analysis is similar in spirit to a combination of singing voice separation and vocal F0 estimation proposed in [21] and in [22]. A key difference is that our method uses robust principal component analysis (RPCA), which is considered to be the state-of-the-art for singing voice separation [18]. As shown in Fig. 1, RPCA is used to extract the singing voice, and then the F0 contour is estimated from the singing voice by finding the optimal path over a F0-saliency spectrogram based on subharmonic summation (SHS). This enables us to improve singing voice separation by combining a time-frequency mask based on RPCA with a mask based on harmonic structures. We use the proposed technique to directly edit vocal F0s in popular-music audio signals.

2. PROPOSED FRAMEWORK

In this section, we explain our proposed framework of mutually dependent vocal F0 estimation and singing voice separation for polyphonic music audio signals. One of our goals is to estimate the vocal F0 at each frame of a target music audio signal. Another is to separate the sung melody from the target signal. Since many promising methods of vocal activity detection (VAD) have already been proposed [10, 23, 24], we do not deal with VAD in this paper.

2.1. Singing voice separation

One of the most promising methods for singing voice separation is to focus on the repeating nature of accompanying sounds [13, 18]. The difference between vocal and accompanying sounds is well characterized in the time-frequency domain. Since the timbres of harmonic instruments, such as pianos and guitars, are consistent for each pitch and the pitches are basically discretized at a semitone level, harmonic spectra having the same shape appear repeatedly in the same musical piece. The spectra of unpitched instruments (*e.g.*, drums) also tend to appear repeatedly. Vocal spectra, in contrast, rarely have the same shape because the timbres and pitches of vocal sounds vary significantly and continuously over time.

In our framework we use robust principal component analysis (RPCA) to separate non-repeating components, as vocal sounds, from a polyphonic spectrogram [18] (see Fig. 2). We decompose an input matrix (spectrogram) M into a low-rank matrix L and a sparse matrix S by solving the following convex optimization problem:

$$\text{minimize} \;\; \|L\|_* + \lambda \|S\|_1 \;\; (\text{subject to} \;\; L + S = M), \qquad (1)$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ represent the nuclear norm and the L1-norm, respectively. λ is a positive parameter that controls the balance between the low-rankness of L and the sparsity of S. To find the optimal L and S, we use an efficient inexact version of the augmented Lagrange multiplier (ALM) algorithm [25].

When RPCA is applied to the spectrogram of a polyphonic music signal, spectral components having repeating structures are allocated to L and the other varying components are allocated to S. We then make a time-frequency binary mask by comparing each element of L with the corresponding element of S. The sung melody is extracted by applying the binary mask to the original spectrogram.

2.2. Vocal F0 estimation

We propose an efficient method that tries to find the optimal F0 path over a saliency spectrogram indicating how likely the vocal F0 is to exist at each time-frequency bin by using the Viterbi algorithm [26]. We test three variants of saliency functions obtained by subharmonic summation (SHS) [27], PreFEst [7], and MELODIA [10].

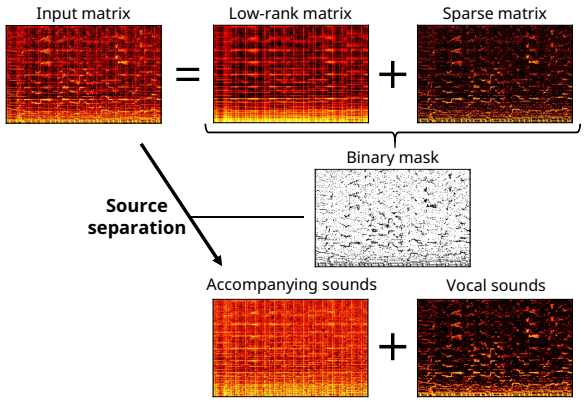

Fig. 2. Singing voice separation based on robust principal component analysis (RPCA).

2.2.1. *Salience functions*

SHS [27] is a standard algorithm that underlies many vocal F0 estimation methods [10, 28]. A salience function $H(t, s)$ is formulated on a logarithmic scale as follows:

$$H(t, s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2 n), \qquad (2)$$

where t and s indicate a frame index and a logarithmic frequency [cents], respectively, $P(t, s)$ represents the power at frame t and frequency s, N is the number of harmonic partials considered, and h_n is a decaying factor (0.86^{n-1} in this paper). The log-frequency power spectrum $P(t, s)$ is calculated from the short-time Fourier transform (STFT) spectrum via spline interpolation. The frequency resolution of $P(t, s)$ is 200 bins per octave (6 cents per bin). Before computing the salience function, we apply to the original spectrum the A-weighting function[1], which takes into account the non-linearity of human auditory perception.

PreFEst [7] is a statistical multipitch analyzer that is considered to be still competitive for vocal F0 estimation. It can be used for computing a salience function. More specifically, an observed spectrum is approximated as a mixture of superimposed harmonic structures. Each harmonic structure is represented as a Gaussian mixture model (GMM) in which each Gaussian corresponds to the energy distribution of a harmonic partial. To learn model parameters, we can use the expectation-maximization (EM) algorithm. The salience function is then obtained as the mixing weights of those harmonic structures. The postprocessing step called PreFEst-*back-end*, which tracks the F0 contour in a multi-agent framework is not used in this paper.

MELODIA [10] is the state-of-the-art method of vocal F0 estimation. It computes a salience function from the spectral peaks of a target music signal after applying an equal-loudness filter. The melody F0 candidates are then selected from the peaks of the salience function and grouped based on time-frequency continuity. Finally, the melody contour is selected from the candidate contours by focusing on the characteristics of vocal F0s. The implementation of MELODIA we use is provided as a vamp plug-in[2].

[1] replaygain.hydrogenaud.ioproposalequal_loudness.html
[2] mtg.upf.edu/technologies/melodia

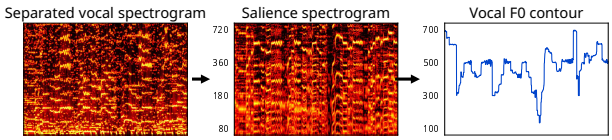

Fig. 3. Vocal F0 estimation based on subharmonic summation (SHS) and Viterbi search

2.2.2. *Viterbi search*

Given a salience function as a time-frequency spectrogram, we estimate the optimal melody contour $\hat{S}$ by solving an optimal path problem formulated as follows:

$$\hat{S} = \underset{s_1,\ldots,s_T}{\text{argmax}} \sum_{t=1}^{T-1} \left\{ \log a_t H(t, s_t) + \log T(s_t, s_{t+1}) \right\}, \quad (3)$$

where $T(s_t, s_{t+1})$ is a transition probability that indicates how likely the current F0 s_t is to move on to the next F0 s_{t+1}, and a_t is a normalization factor that makes the salience values sum to 1 within a range of F0 search. $T(s_t, s_{t+1})$ is given by the Laplace distribution, $\mathcal{L}(s_t - s_{t+1}|0, 150)$, with a zero mean and a standard deviation of 150 cents. The time frame interval is 10 msec. Optimal $\hat{S}$ can be effectively found by using the Viterbi search. Although MELODIA has its own F0 tracking and melody selection algorithm, in this paper we use the Viterbi search for a salience spectrogram obtained by MELODIA in order to purely compare the three salience functions.

2.3. Singing voice separation based on vocal F0s

Assuming that vocal spectra preserve their original harmonic structures and the energy of those spectra is localized on harmonic partials after singing voice separation based on RPCA, we make, in a way similar that of [16], a binary mask M_h that passes only harmonic partials of given vocal F0s:

$$M_h(t, f) = \begin{cases} 1 & \text{if } nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2}, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where F_t is the vocal F0 estimated from frame t, n is the index of a harmonic partial, and w is a frequency width for extracting the energy around each harmonic partial.

We integrate the harmonic mask M_h with the binary mask M_r obtained using the RPCA-based method described in Section 2.1. Finally, a vocal spectrogram P_v and an accompanying spectrogram P_a are given by

$$\begin{aligned} P_v(t, f) &= M_b(t, f) M_h(t, f) P(t, f), \\ P_a(t, f) &= P(t, f) - P_v(t, f), \end{aligned} \quad (5)$$

where P is the original spectrogram of a polyphonic music signal. The separated vocal signals and accompanying signals are obtained by calculating the inverse STFT of P_v and P_a.

3. APPLICATION TO SINGING VOICE EDITING

We use the proposed framework for manipulating vocal F0s included in polyphonic music signals. Our system enables users to add several types of vocal expressions such as vibrato and glissando, to an arbitrary musical note specified on the GUI interface without affecting the timbres of singing voices and accompanying instrument sounds. Example audio files are available on our website.[3]

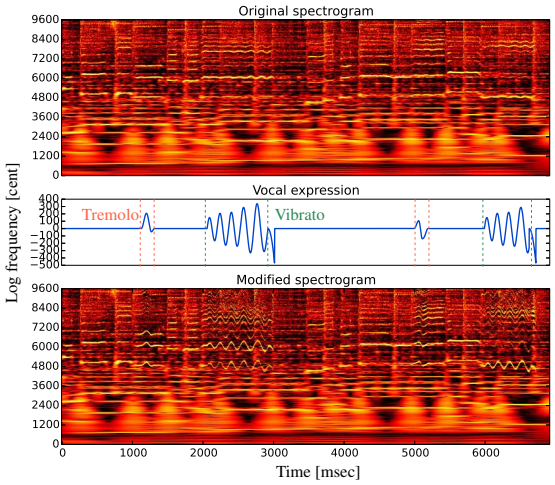

Fig. 4. Example of vocal F0 editing for a piece of popular music (RWC-MDB-P-2001 No.007). From the top to the bottom are shown the original polyphonic spectrogram, the vocal expressions to be attached, and the modified spectrogram.

Here we briefly explain the architecture of the vocal F0 editing system. A target music signal is first converted into a log-frequency amplitude spectrogram by using constant-Q transform [29]. The F0 contour of the singing voice is estimated by using the method described in Section 2.2, and the vocal spectrogram is then separated from the mixture spectrogram by using the method described in Section 2.3. A naive way of changing the F0 of each frame is to just shift the vocal spectrum of each frame along the log frequency axis. That, however, changes the vocal timbre. We therefore first estimate the spectral envelope of the vocal spectrum and then preserve it by modifying the power of each harmonic partial. Finally, a modified music signal is synthesized from the sum of the modified vocal spectra and the separated accompanying spectra by using inverse constant-Q transform [29] with a phase reconstruction method [30].

All these processes are done in the log-frequency domain. This is the first system that applies RPCA to log-frequency spectrograms obtained using a constant-Q transform instead of linear-frequency spectrograms obtained using a short-time Fourier transform (STFT). Figure 4 shows an example of vocal F0 editing, in which vocal expressions such as *vibrato* and *tremolo* are attached to the vocal F0 contour in a polyphonic music signal.

4. EVALUATION

This section describes our experiments evaluating the performances of the proposed singing voice separation and vocal F0 estimation.

4.1. Experimental conditions

The "MIR-1K" dataset[4] and the "RWC Music Database: Popular Music" (RWC-MDB-P-2001) [31] were used in this evaluation. The former contains 110 song clips of 20-110 sec (16 kHz); the latter contains 100 song clips of 125-365 sec (44.1 kHz). The clips of the MIR-1K dataset were with a signal-to-accompaniment ratio of 0

[3] winnie.kuis.kyoto-u.ac.jp/members/ikemiya/demo/icassp2015/

[4] sites.google.com/site/unvoicedsoundseparation/mir-1k

Table 1. Parameter settings.

	window size	interval	N	k	w
MIR-1K	2048	160	10	1.0	80
RWC	4096	441	20	1.0	100

Table 2. Experimental results of vocal F0 estimation. The average accuracy [%] over all clips in each dataset are shown.

MIR-1K (signal-to-accompaniment ratio 0 dB)

Vocal sep.	SHS-V	PreFEst-V	MELODIA-V	MELODIA
None	66.96	50.79	76.88	78.09
RPCA	74.49	56.68	80.59	79.43

RWC-MDB-P-2001

Vocal sep.	SHS-V	PreFEst-V	MELODIA-V	MELODIA
None	71.50	70.07	67.79	69.97
RPCA	77.41	71.01	72.26	69.43

[dB]. The both datasets were used for vocal F0 estimation and only the MIR-1K was used for singing voice separation. The parameters of the STFT (window size and shifting interval [samples]), SHS (the number N of harmonic partials), RPCA (k described in [18]) and the harmonic mask (w [Hz]) are listed in Table 1. The range of the vocal F0 search was set to 80-720 Hz.

4.2. Experimental results of vocal F0 estimation

We tested the following four methods of vocal F0 estimation.

SHS-V: A-weighting function + SHS + Viterbi

PreFEst-V: PreFEst (salience function) + Viterbi

MELODIA-V: MELODIA (salience function) + Viterbi

MELODIA: The original MELODIA algorithm

The raw pitch accuracy (RPA) obtained with and without singing voice separation based on RPCA was measured for each method. The RPA was defined as the ratio of the number of frames in which correct vocal F0s were detected to the total number of voiced frames, and a correct F0 was defined as a detected F0 within 50 cents (*i.e.*, half semitone) of the actual F0. The performance of vocal activity detection (VAD) was not measured in this study.

As seen in Table 2, the experimental results showed that the proposed method SHS-V performed well with both datasets. We found that singing voice separation was a great help, especially with SHS-V that is a simple SHS-based method. PreFEst-V did not work well with the MIR-1K dataset because many clips in that dataset contained melodic instrumental sounds with salient harmonic structure (*e.g.*, a piano and strings along with a singing voice).

4.3. Experimental results of singing voice separation

We tested the following four methods of singing voice separation.

RPCA: Using only RPCA mask [18]

RPCA-F0: Using RPCA mask + harmonic mask (proposed)

RPCA-F0-GT: Using RPCA mask + harmonic mask (made by using ground-truth F0s)

IDEAL: Using ideal binary mask (upper bound)

In this experiment we used the SHS-V method for vocal F0 estimation because its overall performance was better than that of the other methods. The BSS-EVAL toolkit [32] was used for evaluating the quality of separated audio signals in terms of source-to-interference ratio (SIR), sources-to-artifacts ratio (SAR), and source-to-distortion ratio (SDR) by comparing separated vocal sounds with ground-truth isolated vocal sounds. Normalized SDR (NSDR) [18] was also calculated for evaluating the improvement of the SDR from that of the original music signals. The final scores, GSIR, GSAR, GSDR and GNSDR were obtained by taking the averages over all 110 clips of MIR-1K, weighted by their lengths. Since this paper does not deal with VAD and intended to examine the effect of harmonics mask for singing voice separation, we used only voiced frames for evaluation, *i.e.*, the amplitudes of separated signals in unvoiced frames were set to 0 when computing the evaluation scores.

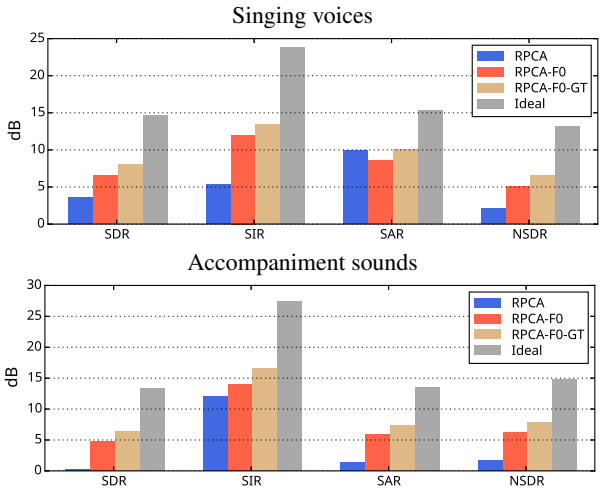

Fig. 5. Experimental results of singing voice separation for the MIR-1K dataset: Source separation quality for singing voices (top) and accompanying sounds (bottom)

The experimental results showed that, by all measures except GSAR, the proposed RPCA-F0 method worked better than the RPCA (Fig. 5). Although vocal F0 estimation often failed, removing the spectral components of non-repeating instruments (*e.g.*, a bass guitar) significantly improved the separation of both vocal and accompanying signals. The proposed method outperformed the state-of-the-art methods in the Music Information Retrieval Evaluation eXchange (MIREX 2014)[5].

5. CONCLUSION

This paper proposed a novel framework for improving both vocal F0 estimation and singing voice separation by making effective use of the mutual dependency of those tasks. In the first step, we perform blind singing voice separation without assuming singing voices to have harmonic structures by using robust principal component analysis (RPCA). In the second step, we detect the vocal contour in the separated vocal spectrogram by using a simple saliency-based method called subharmonic summation. In the last step, we accurately extract the singing voice by making a binary mask based on vocal harmonic structures and the RPCA results. These techniques enable users to freely edit vocal F0s in music signals in existing CD recordings for active music listening. In the future we plan to integrate both tasks into a unified probabilistic model jointly optimizing their results in a principled manner.

[5] www.music-ir.org/mirex/wiki/2014:Singing_Voice_Separation_Results

6. REFERENCES

[1] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. ICASSP*, 2007, pp. 1441–1444.

[2] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Drumix: An audio player with real-time drum-part rearrangement functions for active music listening," in *IPSJ Journal*, 2007, vol. 48, pp. 1229–1239.

[3] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. ICASSP*, 2013, pp. 888–891.

[4] N. J. Bryan, G. J. Mysore, and G. Wang, "Source separation of polyphonic music with interactive user-feedback on a piano roll display," in *Proc. ISMIR*, 2013, pp. 119–124.

[5] Y. Ohishi, D. Mochihashi, H. Kameoka, and K. Kashino, "Mixture of gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations," in *Proc. ICASSP*, 2014, pp. 3714–3718.

[6] H. Fujihara and M. Goto, "Concurrent estimation of singing voice F0 and phonemes by using spectral envelopes estimated from polyphonic music," in *Proc. ICASSP*, 2011, pp. 365–368.

[7] M. Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," in *Speech Communication*, 2004, vol. 43, pp. 311–329.

[8] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," in *IEEE Trans. on Audio, Speech and Language Processing*, 2010, vol. 18, pp. 2145–2154.

[9] K. Dressler, "An auditory streaming approach for melody extraction from polyphonic music," in *Proc. ISMIR*, 2011, pp. 19–24.

[10] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," in *IEEE Trans. on Audio, Speech and Language Processing*, 2012, vol. 20, pp. 1759–1770.

[11] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," in *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 2014, pp. 228–237.

[12] C. L. Hsu and J. R. Jang, "Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion," in *Proc. ISMIR*, 2010, pp. 525–530.

[13] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," in *IEEE Trans. on Audio, Speech and Language Processing*, 2013, vol. 21, pp. 71–82.

[14] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," in *IEEE Signal Process. Mag.*, 2014, vol. 31, pp. 118–134.

[15] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," in *IEEE Trans. on Audio, Speech and Language Processing*, 2007, vol. 15, pp. 1475–1487.

[16] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, 2008.

[17] E. Cano, C. Dittmar, and G. Schuller, "Efficient implementation of a system for solo and accompaniment separation in polyphonic music," in *Proc. EUSIPCO*, 2012, pp. 285–289.

[18] P. S. Huang, S. Deeann Chen, P. Smaragdis, and M. H. Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP*, 2012, pp. 57–60.

[19] D. Fitzgerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," in *ISAST Trans. on Electronic and Signal Processing*, 2010, vol. 4, pp. 62–73.

[20] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," in *IEEE J. Selected Topics in Signal Processing*, 2011, vol. 5, pp. 1180–1191.

[21] C. L. Hsu, D. Wang, J. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," in *IEEE Trans. on Audio, Speech and Language Processing*, 2012, vol. 20, pp. 1482–1491.

[22] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," in *IEEE Trans. on Audio, Speech and Language Processing*, 2014, vol. 22, pp. 1884–1893.

[23] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Proc. ICASSP*, 2008, pp. 1885–1888.

[24] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics," in *IEEE Journal of Selected Topics in Signal Processing*, 2011, vol. 5, pp. 1252–1261.

[25] Y. Ma Z. Lin, M. Chen, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," in *Mathematical Programming*, 2009.

[26] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *IEEE*, 1989, vol. 77, pp. 257–286.

[27] D. J. Hermes, "Measurement of pitch by subharmonic summation," in *J. Acoust. Soc. Am.*, 1988, vol. 83, pp. 257–264.

[28] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing melody extraction in polyphonic music by harmonic tracking," in *Proc. ISMIR*, 2007, pp. 373–374.

[29] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. SMC*, 2010.

[30] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," in *IEEE Trans. on Signal Proc.*, 1993, vol. 41, pp. 3549–3554.

[31] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," in *IEEE Trans. on Audio, Speech and Language Processing*, 2006, vol. 14, pp. 1462–1469.