VOCAL ACTIVITY INFORMED SINGING VOICE SEPARATION WITH THE IKALA DATASET

Tak-Shing Chan[1]*, Tzu-Chun Yeh*[2]*, Zhe-Cheng Fan*[2]*, Hung-Wei Chen*[3]*, Li Su*[1]*, Yi-Hsuan Yang*[1]*, Roger Jang*[2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[3] iKala Interactive Media Inc., Taiwan

ABSTRACT

A new algorithm is proposed for robust principal component analysis with predefined sparsity patterns. The algorithm is then applied to separate the singing voice from the instrumental accompaniment using vocal activity information. To evaluate its performance, we construct a new publicly available iKala dataset that features longer durations and higher quality than the existing MIR-1K dataset for singing voice separation. Part of it will be used in the MIREX Singing Voice Separation task. Experimental results on both the MIR-1K dataset and the new iKala dataset confirmed that the more informed the algorithm is, the better the separation results are.

Index Terms— Low-rank and sparse decomposition, singing voice separation, informed source separation

1. INTRODUCTION

The robust principal component analysis (RPCA) algorithm decomposes an input matrix (e.g., a spectrogram) $X \in \mathbb{R}^{m \times n}$ into a low-rank matrix $A \in \mathbb{R}^{m \times n}$ plus a sparse matrix $X - A$. The problem can be formulated as [1]

$$\min_{A} \|A\|_* + \lambda \|X - A\|_1 \,, \tag{1}$$

where $\|A\|_* = \text{tr}(\sqrt{A^T A})$ is the trace norm of A, $\|\cdot\|_1$ denotes the entrywise l_1-norm, and λ is a positive constant which can be set to $1/\sqrt{\max(m,n)}$ [2]. The trace-norm relaxation of the rank permits efficient convex minimization [1]. Unlike the classical PCA, RPCA is robust against outliers.

The singing voice separation problem, which aims to separate the singing voice and instrumental accompaniment from monaural mixtures, can be applied to singing voice analysis [3, 4], beat tracking [5], instrument detection [6], karaoke applications [7], and so on. For singing voice separation, Rafii and Pardo's [8] REPET is the first to assume sparse voice plus repeating background, whereas Huang *et al.*'s [9] RPCA approach contains the first explicit model of sparse voice plus low-rank music. Both exploit redundancy and repetitive patterns at longer time scales [10–12] instead of separating frame by frame. Please refer to [13] for a recent review of other competing methods such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA).

At least two trends emerge in recent years. One is the combination of existing building blocks in new ways [13, 14]. The other is informed source separation [15–19]. The latter approach is receiving increasing attention as the number of musicians usually exceeds the number of microphones, rendering blind source separation ill-posed. Furthermore, music sources are particularly challenging to separate because they are typically highly correlated in time and frequency [20]. Side information is necessary for better results [10].

Although many score-informed source separation algorithms have been proposed for music signals, for popular music the musical scores may not be as available [20]. Recently we have observed that the sparse matrix from RPCA often contains the predominant instrument or percussion and "this issue might be partially solved by performing vocal detection first to exclude non-vocal segments" [6]. However, RPCA and REPET do not explicitly capitalize on the fact that a popular song is composed of vocal (with both singing and instrumental accompaniment) and non-vocal (with only accompaniment) segments, which is an important information source.

This paper presents a modifed RPCA algorithm to explore this research direction. The algorithm exploits the vocal/non-vocal patterns of the audio clip to coerce parts of the spectrogram to be non-vocal. While vocal activity detection has been studied extensively [21, 22], to the best of our knowledge, this work represents one of the first attempts to incorporate vocal activity information into the RPCA algorithm. Though Bryan *et al.*'s [19] Interactive Source Separation Editor (ISSE) is earlier, it modifies PLCA and not RPCA.

To further advance research on singing voice separation, we present a new iKala dataset that has a higher quality than the existing MIR-1K dataset [23]. While the audio clips in MIR-1K are usually shorter than 10 seconds and therefore lacking the non-vocal regions in popular songs, the iKala dataset is composed of 352 30-second clips featuring longer

This work was supported by the Academia Career Development Program. We are sincerely grateful to industrial partner iKala Interactive Media Inc. for sharing the dataset with the research community.

instrumental solos. Part of the dataset is reserved for the Singing Voice Separation task of the Music Information Retrieval Evaluation eXchange (MIREX).[1] The remaining audio clips and associated data will be made publicly available to the research community for academic purposes through a webpage.[2] This study experiments with three different vocal activity informed scenarios: no masks, vocal/non-vocal masks, and ideal binary masks.

In what follows, we describe the proposed algorithm in Section 2 and the new dataset in Section 3. Then, we present experimental results in Section 4 and conclude in Section 5.

Algorithm 1 RPCA with Predefined Sparsity Patterns

Input: $X \in \mathbb{R}^{m\times n}, M \in \{0,1\}^{m\times n}, \lambda \in \mathbb{R}, \mu \in \mathbb{R}^{\infty}$
Output: A_k, E_k
1: Let $E_1 = 0, Y_1 = X/\max\left(\|X\|_2, \lambda^{-1}\|X\|_\infty\right), k = 1$
2: **while** not converged **do**
3: $A_{k+1} \leftarrow \arg\min_A \mathcal{L}(A, E_k, Y_k; \mu_k)$
4: $E_{k+1} \leftarrow (\mathbf{1} - M) \circ \arg\min_E \mathcal{L}(A_{k+1}, E, Y_k; \mu_k)$
5: $Y_{k+1} \leftarrow Y_k + \mu_k(X - A_{k+1} - E_{k+1})$
6: $k \leftarrow k + 1$
7: **end while**

2. ROBUST PRINCIPAL COMPONENT ANALYSIS WITH PREDEFINED SPARSITY PATTERNS (RPCA*s*)

By adding a simple constraint to (1), we can coerce parts of the input matrix to be non-sparse (or non-vocal, in terms of singing voice separation). Given an input matrix $X \in \mathbb{R}^{m\times n}$ and a predefined sparsity mask $M \in \{0,1\}^{m\times n}$, RPCA*s* solves the following optimization problem:

$$\min_{A,E} \|A\|_* + \lambda\|E\|_1 \text{ s.t. } X = A + E \text{ and } M \circ E = 0\,, \quad (2)$$

where $\circ$ denotes the Hadamard product. This problem can be solved by the inexact ALM method [24], which minimizes the partial augmented Lagrangian function [1, 24]

$$\mathcal{L}(A, E, Y; \mu) = \|A\|_* + \lambda\|E\|_1 + \text{tr}(Y^T(X - A - E)) + \frac{\mu}{2}\|X - A - E\|_F^2\,, \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm and μ is a positive penalty parameter. This is done by alternately solving for A and E via the singular value shrinkage operator [2], then updating the Lagrangian multiplier Y. The additional constraint $M \circ E = 0$ should be enforced at each iteration. This is summarized in Algorithm 1. Note that in terms of the constraints we could also interpret the current problem as half RPCA and half matrix completion [1]. If M is all zeros, this formulation coincides with the original RPCA with no masks.

[1] http://www.music-ir.org/mirex/wiki/2014:Singing_Voice_Separation

[2] http://mac.citi.sinica.edu.tw/ikala/

In addition to the all-zero mask, two more sparsity patterns are proposed in this work. First, we can use the *vocal/non-vocal mask* defined by:

$$M_{ij} = \begin{cases} 1, & \text{if } P_j = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where P denotes the human annotations of vocal pitch and 0 means non-vocal [23]. As will be described in the next section, both the MIR-1K and the iKala datasets come with pitch contour annotations, so we can use them to obtain the vocal/non-vocal masks. But state-of-the-art vocal detection methods have also been shown effective [21, 22] and may be readily applicable. We leave this for future work.

Second, we can use the *ideal time-frequency binary mask* defined as follows [25]:

$$M^*_{ij} = \begin{cases} 1, & \text{if } |K_{ij}|^2 > |V_{ij}|^2, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where K is the spectrogram of the ground truth music accompaniment and V is the spectrogram of the ground truth singing voice. These masks are quite informative (especially the last one), so we can expect increased performance when we use them. In our study, we consider the ideal binary mask as an oracle method to test the best possible performance obtainable by RPCA*s*. In addition, the ideal binary mask can be used without RPCA*s* to get the best possible performance for all possible algorithms. In practice, one might employ the classification-based source separation approach proposed by Wang *et al.* [26] to estimate the binary mask from audio.

3. THE IKALA DATASET

The MIR-1K dataset is the first public dataset specifically created for singing voice separation [23]. Three criteria are considered while compiling the dataset: 1) voice and music recorded separately, 2) comprehensive manual annotations, and 3) dataset publicly available. It consists of 1,000 clips of Chinese popular songs recorded at 16kHz, with clean music accompaniments retrieved from Karaoke songs and singing voices sung by amateur singers [23]. It also comes with human-labeled pitch values, vocal/non-vocal regions, and lyrics, among others. While MIR-1K has been very useful in the past, rapid advances in computing technology demand even higher quality datasets. This is where iKala comes in.

The iKala dataset contains 352 30-second clips of Chinese popular songs in CD quality. The singers and musicians are professionals. The dataset is human-labeled for pitch contours and timestamped lyrics. Moreover, as the clips are longer, the iKala dataset contains non-vocal regions (e.g., instrumental solos) that may challenge separation algorithms that assume the presence of human voice throughout the audio clip. Please see Table 1 for a detailed comparison between

	MIR-1K [23]	iKala
Sampling rate	16kHz	44.1kHz
Singer quality	Amateur	Professional
Clip duration	4-13s	30s
Number of clips	1,000	252+100
Voice recorded separately	Yes	Yes
Pitch contour annotations	Yes	Yes
Voice type annotations	Yes	No
Lyrics with speech	Yes	No
Lyrics with timestamps	No	Yes
Separate chorus and verse	No	Yes
Instrumental solo	No	Yes

Table 1. Comparison of MIR-1K and iKala datasets

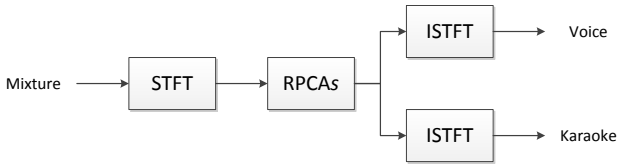

Fig. 1. Block diagram of our proposed system.

the two datasets. Following the convention of MIREX, 100 clips from iKala are reserved for the annual MIREX Singing Voice Separation task and will not be made public. However, researchers will be able to use the 252-clip public subset for parameter fine-tuning or for developing supervised or semi-supervised algorithms for source separation [12, 27]. There are no overlapping songs between MIR-1K and iKala.

4. EXPERIMENTS

We will study the effect of the following three levels of informedness on separation performance:

- **RPCA** with no masks (least informed).
- **RPCA*s*** with vocal/non-vocal masks (informed).
- **RPCA*s*-IBM** with ideal binary masks (most informed).

In addition, we will use **REPET-SIM** [28] with the default parameters as a baseline, and **IBM** without RPCA*s* as the theoretical upper bound. Our evaluation will run on both the iKala and the MIR-1K datasets. Voice and music will be mixed at 0 dB. To reduce hard disk usage, all iKala clips will be downsampled to 22,050 Hz. Please refer to our main setup in Fig. 1. For each clip in the iKala dataset (excluding the MIREX-reserved set), we will use a short-time Fourier transform (STFT) with a 1,411-point Hann window, and for each clip in MIR-1K, we will use STFT with a 1,024-point Hann window, both with 75% overlap, following [9]. The spectrograms thus obtained will contain a magnitude part and a phase part P. The magnitude part will further be decomposed by RPCA*s* into voice and music components which will then be separately reconstructed in the time domain via inverse STFT using the original phase P [29], a common practice for source separation in the spectrogram domain [10]. Although it is possible to perform source separation in the time domain and thereby better exploits phase information [30, 31], this is beyond the scope of this work.

(a) iKala, voice

Method	GNSDR	GSIR	GSAR
REPET-SIM	3.09	7.25	10.0
RPCA	2.16	5.66	10.8
RPCA*s*	4.42	10.7	9.21†
RPCA*s*-IBM	8.13	26.0	9.31†
IBM	12.3	23.7	14.1

(b) iKala, music

Method	GNSDR	GSIR	GSAR
REPET-SIM	5.24	5.77	8.15
RPCA	4.52	5.48	6.12
RPCA*s*	6.17	7.07	7.53
RPCA*s*-IBM	8.32	8.45	10.9
IBM	15.9	29.6	12.1

(c) MIR-1K, voice

Method	GNSDR	GSIR	GSAR
REPET-SIM	2.59	4.77	8.81
RPCA	3.20	4.72	10.4
RPCA*s*	5.03	7.80	9.59†
RPCA*s*-IBM	10.2	22.4	10.7
IBM	13.5	21.3	14.5

(d) MIR-1K, music

Method	GNSDR	GSIR	GSAR
REPET-SIM	2.83	4.55	9.82
RPCA	3.32	5.41	9.20
RPCA*s*	4.52	6.48	10.4
RPCA*s*-IBM	6.33	7.23	15.1
IBM	13.2	24.0	13.7

Table 2. Separation quality (in dB) for the voice and music channels for the (a)(b) iKala and (c)(d) MIR-1K datasets.

The quality of separation is assessed in terms of source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) [32], which are computed for the vocal part v and the instrumental part a, respectively. We computed these ratios by using BSS Eval Version 3.0 [32]. We compute the normalized SDR (NSDR) by $\text{SDR}(\hat{v}, v) - \text{SDR}(x, v)$. Moreover, we aggregate the performance over all the test clips by taking the weighted average, with weight proportional to the length of each clip [23]. The resulting measures are denoted as GNSDR, GSIR, and GSAR, respectively. Note the later two are not normalized.

Results for the iKala and the MIR-1K datasets are shown in Table 2(a)(b) and Table 2(c)(d), respectively. We ascertain

the effect of informedness with 12 one-tailed paired t-tests, confirming that RPCAs-IBM $>$ RPCAs $>$ RPCA (all with $p < 10^{-10}$, except the three denoted by daggers in Table 2, which are not significant). We observe that RPCA performs slightly better than REPET-SIM for MIR-1K. However, the clips in MIR-1K are too short and cannot reflect the prevalent non-vocal regions in full songs, a case for which RPCA falls short of. Clearly, REPET-SIM outperforms RPCA for the iKala dataset. By taking into account voice activity information, the proposed RPCAs algorithms lead to much better performance than RPCA and REPET-SIM in terms of GNSDR and GSIR, especially for vocal GSIR. However, the RPCAs algorithms perform slightly unfavorably for GSAR, except for the RPCAs-IBM method. This shows a possible limitation of the vocal/non-vocal mask. Interestingly, for both vocal GSIR and music GSAR from MIR-1K, RPCAs-IBM performs better than IBM, suggesting that RPCAs might be trading music GSAR for vocal GSIR.

For the iKala dataset, we also show the boxplots for NSDR, SIR and SAR for the singing voice in Fig. 2. It can be seen that RPCAs outperforms the competing methods REPET-SIM and RPCA in NSDR and SIR.

5. CONCLUSIONS

In this paper, we have proposed a novel vocal activity informed singing voice separation algorithm and presented the new iKala dataset. Evaluation on the public 252-clip subset as well as the existing MIR-1K dataset showed that vocal/non-vocal information would significantly increase performance, except for SAR. Although we assume vocal/non-vocal information is known, in practice we can use an automatic vocal activity detector [21] to obtain this side information. In the future, we expect iKala's pitch contour information together with time-stamped lyrics would further enhance the quality of separation. We also look forward to more research on this problem due to the organization of source separation evaluation campaigns such as MIREX and SiSEC [33].

6. REFERENCES

[1] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, 2009.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[3] T. Nakano, K. Yoshii, and M. Goto, "Vocal timbre analysis using latent dirichlet allocation and cross-gender vocal timbre similarity," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 5202–5206.

[4] C.-Y. Sha, Y.-H. Yang, Y.-C. Lin, and H. H. Chen, "Singing voice timbre classification of chinese popular music," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013.

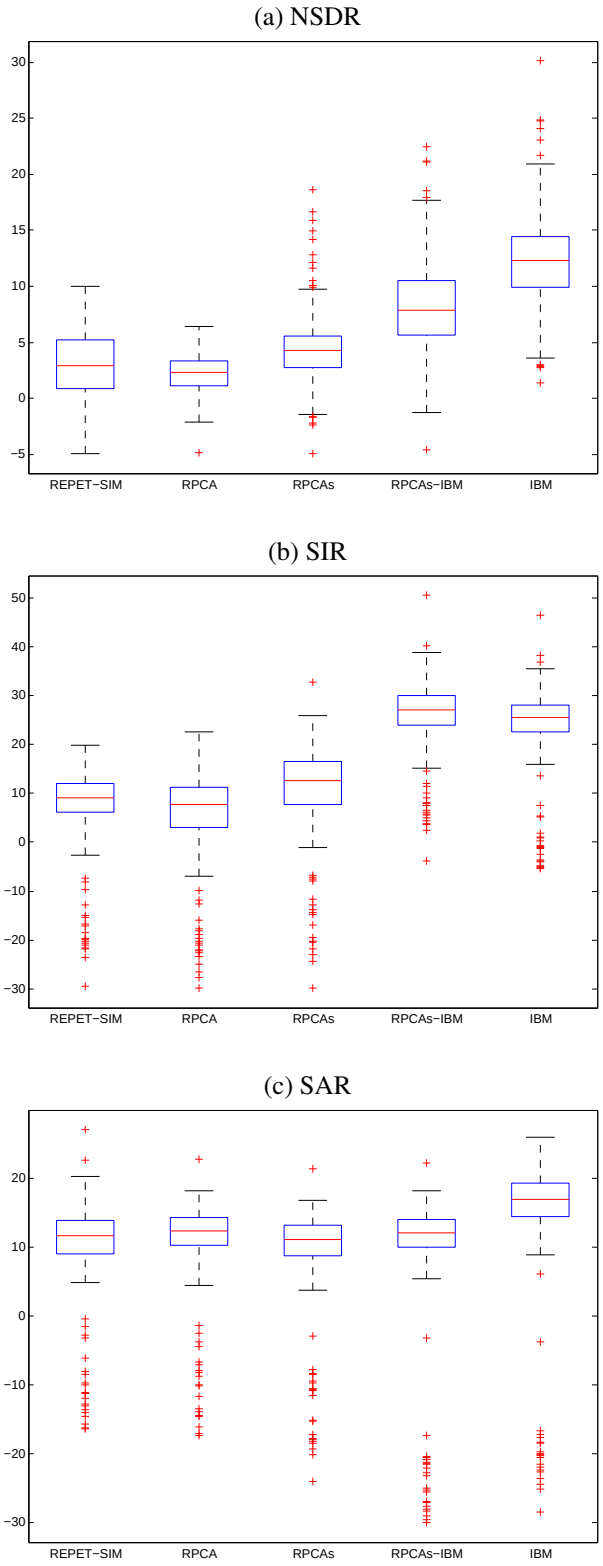

Fig. 2. Boxplots of (a) normalized source-to-distortion ratio (NSDR), (b) source-to-interference ratio (SIR), and (c) source-to-artifact ratio (SAR) for the voice part of the iKala dataset.

[5] J. R. Zapata and E. Gomez, "Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 51–55.

[6] L. Su and Y.-H. Yang, "Sparse modeling for artist identification: Exploiting phase information and vocal separation," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2013, pp. 349–354.

[7] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2012.

[8] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the underlying repeating structure," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2011, pp. 221–224.

[9] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 57–60.

[10] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[11] P. Sprechmann, Alexander M. B., and Guillermo S., ," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2012, pp. 67–72.

[12] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2013.

[13] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1884–1893, 2014.

[14] 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," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228–237, 2014.

[15] A Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services*, 2013, pp. 1–4.

[16] Z. Duan and B. Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE J. Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[17] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 888–891.

[18] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2012.

[19] 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. Int. Soc. Music Info. Retrieval Conf.*, 2013, pp. 119–124.

[20] S. Ewert, B. Pardo, M. Muller, and M.D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[21] B. Lehner, G. Widmer, and R. Sonnleitner, "On the reduction of false positives in singing voice detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 7480–7484.

[22] K. Yoshii, H. Fujihara, T. Nakano, and M. Goto, "Cultivating vocal activity detection for music audio signals in a circulation-type crowdsourcing ecosystem," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 624–628.

[23] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. Audio, Speech & Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[24] D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Method*, Academic Press, 1982.

[25] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Pierre Divenyi, Ed., pp. 181–197. Springer US, 2005.

[26] Y. Wang, A Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[27] F. Weninger, J. Feliu, and B. Schuller, "Supervised and semi-supervised suppression of background music in monaural speech recordings," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 61–64.

[28] Z. Rafii and B. Pardo, "Music-voice separation using the similarity matrix," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2012, pp. 583–588.

[29] D. Ellis, "A phase vocoder in Matlab," 2002, [Online] http://www.ee.columbia.edu/ dpwe/resources/matlab/pvoc/.

[30] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Beyond NMF: time-domain audio source separation without phase reconstruction," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2013, pp. 369–374.

[31] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 7475–7479.

[32] E. Vincent, R. Gribonval, and C. Fèvotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech & Language Processing*, vol. 16, no. 4, pp. 766–778, 2008.

[33] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, 2012, pp. 414–422.