Scientific Research Publishing

Advances in Pure Mathematics, 2018, 8, 419-427
http://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Low-Rank Sparse Representation with Pre-Learned Dictionaries and Side Information for Singing Voice Separation

Chenghong Yang, Hongjuan Zhang

Department of Mathematics, Shanghai University, Shanghai, China
Email: Ychenghong@i.shu.edu.cn

How to cite this paper: Yang, C.H. and Zhang, H.J. (2018) Low-Rank Sparse Representation with Pre-Learned Dictionaries and Side Information for Singing Voice Separation. *Advances in Pure Mathematics*, **8**, 419-427.
https://doi.org/10.4236/apm.2018.84024

Received: March 21, 2018
Accepted: April 21, 2018
Published: April 24, 2018

Abstract

At present, although the human speech separation has achieved fruitful results, it is not ideal for the separation of singing and accompaniment. Based on low-rank and sparse optimization theory, in this paper, we propose a new singing voice separation algorithm called Low-rank, Sparse Representation with pre-learned dictionaries and side Information (LSRi). The algorithm incorporates both the vocal and instrumental spectrograms as sparse matrix and low-rank matrix, meanwhile combines pre-learning dictionary and the reconstructed voice spectrogram form the annotation. Evaluations on the iKala dataset show that the proposed methods are effective and efficient for singing voice separation.

Keywords

Singing Voice Separation, Low-Rank and Sparse, Dictionary Learning

1. Introduction

Separating singing voice from music recording is very useful in many applications, such as music information retrieval, singer identification and lyrics recognition and alignment [1]. Although the human auditory system can easily distinguish the vocal and instrumental of music recording, it is extremely difficult for computer systems. In this context, researchers are increasingly concerned with the mining of music information. Many algorithms have been proposed to separate singing voice from music recording.

Robust Principal Component Analysis (RPCA) is a matrix factorization algorithm for solving underlying low-rank and sparse matrices [2]. Suppose we are given a large data matrix M, and know that it may be decomposed as $X = A + E$,

where A is a low-rank matrix and E is a sparse matrix. Based on RPCA, Huang *et al.* [3] have separated singing-voice from music accompaniment. They assumed that the repetitive music accompaniment lies in a low-rank subspace, while the singing voices can be regarded as sparse within songs. The main drawback to this approach is that it is completely unsupervised, just based on the particular properties of each individual components to guide the decomposition. After, Yu *et al.* [4] utilized any pre-learned information and pre-learned universal voice and music dictionaries from isolated singing voice and background music training data. They proposed Low-rank and Sparse representation with Pre-learned Dictionaries (LSPD) for singing voice separation. Chan *et al.* [5] proposed a modified RPCA algorithm. This work represented one of the first attempts to incorporate vocal activity information into the RPCA algorithm, then the vocal activity detection was widely studied [6] [7]. Chan *et al.* [8] proposed to separate singing voice by group-sparse representation with the idea of pitch annotations separation.

In this paper, we present a model named Low-rank, Sparse representation with pre-learned dictionaries and side information (LSRi) under the ADMM framework. First, we pre-learn voice and music dictionaries from isolated singing voice and background music training data, respectively. Then, we use a sparse spectrogram and a low-rank spectrogram to model the singing voice and the background music, respectively. Outside, a residual term is added to capture the components that are not well modeled by either the sparse or the low-rank term. Finally, we combine the reconstructed voice spectrogram from the vocal annotation. Evaluations on the iKala dataset [9] show its better performance than comparison methods.

The rest of this paper is organized as follows. The overview of the music analysis model is presented in Section 2. The description of theoretical knowledge and experimental results are presented in Section 3. Final Section concludes this work.

2. The Proposed Method

Before we come up with our method, let's review the Low-rank and Sparse representation with Pre-learned Dictionaries (LSPD) method [4],

$$\begin{aligned} &\min_{Z_1,Z_2} \|Z_1\|_* + \lambda_1 \|Z_2\|_1 + \lambda_2 \|E\|_1 \\ &s.t.\ X = D_1 Z_1 + D_2 Z_2 + E \end{aligned} \tag{1}$$

where X is the input spectrogram, $D_1 \in R^{m \times k_1}$ is a pre-learned dictionary of the music accompaniment, $D_2 \in R^{m \times k_2}$ is a pre-learned dictionary of the singing voice, $D_1 Z_1$ is the separated instrumentals, $D_2 Z_2$ is the separated voice. E denotes the residual part. λ_1, λ_2 are two weighting parameters for balancing the different regularization terms in this model.

Compared with the unsupervised RPCA algorithm, the LSPD algorithm adds pre-learning dictionary information and improves the separation quality. To

further improve the separation quality of singing voice and music accompaniment, we proposed Low-rank, Sparse Representation with pre-learned dictionaries and side Information (LSRi).

In our model, we considered more prior information *i.e.*, the reconstructed voice spectrogram from the annotation. Model as follows,

$$\begin{aligned} &\min_{Z_1,Z_2} \|Z_1\|_* + \lambda_1 \|Z_2\|_1 + \lambda_2 \|E\|_1 + \frac{\gamma}{2}\|D_2 Z_2 - E_0\|_F^2 \\ &s.t.\ \ X = D_1 Z_1 + D_2 Z_2 + E \end{aligned} \tag{2}$$

Here all parameters in model 2 are in accordance with model 1, and E_0 denotes the reconstructed voice spectrogram from the annotation. $\|\cdot\|_F$ denotes the Frobenius norm. In the following, we also use the ADMM algorithm [10] to solve the optimization problem, by introducing two auxiliary variables J_1 and J_2 as well as three equality constraints,

$$\begin{aligned} &\min_{Z_1,Z_2,J_1,J_2} \|J_1\|_* + \lambda_1 \|J_2\|_1 + \lambda_2 \|E\|_1 + \frac{\gamma}{2}\|D_2 Z_2 - E_0\|_F^2 \\ &s.t.\ \ X = D_1 Z_1 + D_2 Z_2 + E, Z_1 = J_1, Z_2 = J_2 \end{aligned} \tag{3}$$

The unconstrained augmented Lagrangian $\mathcal{L}$ is given by

$$\begin{aligned} \mathcal{L} = &\|J_1^{\mathrm{T}}\|_* + \lambda_1 \|J_2\|_1 + \lambda_2 \|E\|_1 + \frac{\gamma}{2}\|D_2 Z_2 - E_0\|_F^2 \\ &+ \langle Y_1, X - D_1 Z_1 - D_2 Z_2 - E\rangle + \langle Y_2, Z_1 - J_1\rangle + \langle Y_3, Z_2 - J_2\rangle \\ &+ \frac{\mu}{2}\left(\|X - D_1 Z_1 - D_2 Z_2 - E\|_F^2 + \|Z_1 - J_1\|_F^2 + \|Z_2 - J_2\|_F^2\right) \end{aligned} \tag{4}$$

where Y_1, Y_2, Y_3 are the Lagrange multipliers. We then iteratively update the solutions for J_1, Z_1, J_2 and Z_2.

1) Update J_1:

$$J_1 = \arg\min_{J_1} \|J_1\|_* + \frac{\mu}{2}\left\|J_1 - \left(Z_1 + \mu^{-1} Y_2\right)\right\|_F^2 = US_{\frac{1}{\mu}}[\Sigma]V^{\mathrm{T}} \tag{5}$$

where $U\Sigma V = svd\left(Z_1 + \mu^{-1} Y_2\right)$.

2) Update Z_1:

$$\frac{\partial \mathcal{L}}{\partial Z_1} = -D_1^{\mathrm{T}} Y_1 + Y_2 - \mu D_1^{\mathrm{T}}\left(X - D_1 Z_1 - D_2 Z_2 - E\right) + \mu\left(Z_1 - J_1\right) \tag{6}$$

setting $\frac{\partial \mathcal{L}}{\partial Z_1} = 0$, we have

$$Z_1 = \left(D_1^{\mathrm{T}} D_1 + I\right)^{-1}\left(D_1^{\mathrm{T}}\left(X - D_2 Z_2 - E + \mu^{-1} Y_1\right) - \mu^{-1} Y_2 + J_1\right) \tag{7}$$

3) Update J_2:

$$J_2 = \arg\min_{J_2} \lambda_1 \|J_2\|_1 + \frac{\mu}{2}\left\|J_1 - \left(Z_1 + \mu^{-1} Y_3\right)\right\|_F^2 \tag{8}$$

that can be solve by the soft-threshold operator

$$J_2 = S_{\frac{\lambda_1}{\mu}}\left(Z_2 + \mu^{-1} Y_3\right) \tag{9}$$

since the spectrogram is non-negative

$$J_2 = \max\left\{S_{\frac{\lambda_1}{\mu}}\left(Z_2 + \mu^{-1}Y_3\right), 0\right\} \tag{10}$$

where 0 is an all zero matrix of the size as J_2.

4) Update Z_2:

$$\frac{\partial \mathcal{L}}{\partial Z_2} = \gamma D_2^{\mathrm{T}}\left(D_2 Z_2 - E_0\right) - D_2^{\mathrm{T}}Y_1 + Y_3 - \mu D_2^{\mathrm{T}}\left(X - D_1 Z_1 - D_2 Z_2 - E\right) + \mu\left(Z_2 - J_2\right) \tag{11}$$

setting $\frac{\partial \mathcal{L}}{\partial Z_2} = 0$, we have

$$\begin{aligned} Z_2 &= \left(\left(\gamma+\mu\right)D_2^{\mathrm{T}}D_2 + \mu I\right)^{-1}\left(\gamma D_2^{\mathrm{T}}E_0 + D_2^{\mathrm{T}}Y_1 - Y_3 + \mu D_2^{\mathrm{T}}\left(X - D_1 Z_1 - E\right) + \mu J_2\right) \\ &= \left(\left(\frac{\gamma}{\mu}+1\right)D_2^{\mathrm{T}}D_2 + I\right)^{-1}\left(D_2^{\mathrm{T}}\left(X - D_1 Z_1 - E + \frac{\gamma}{\mu}E_0 + \frac{1}{\mu}Y_1\right) - \frac{1}{\mu}Y_3 + J_2\right) \end{aligned} \tag{12}$$

5) Update E:

$$E = \arg\min_E \lambda_2 \|E\|_1 + \frac{\mu}{2}\left\|E - \left(X - D_1 Z_1 - D_2 Z_2 + \mu^{-1}Y_1\right)\right\|_F^2 \tag{13}$$

Similar to J_2,

$$E = \max\left\{S_{\frac{\lambda_2}{\mu}}\left(X - D_1 Z_1 - D_2 Z_2 + \mu^{-1}Y_1\right), 0\right\} \tag{14}$$

Finally, we update the Lagrange multipliers as in [11].

3. Experiment

3.1. Dataset

Our experiment was conducted on the iKala dataset [9]. The iKala dataset contains 252 30-second clips of Chinese popular songs in CD quality. In the following experiments, we randomly select 44 songs for training (*i.e.*, learning the dictionaries D_1 and D_2), leaving 208 songs for testing the performance of separation. To reduce the computational cost and the memory footprint of the proposed algorithm, we down sample all the audio recordings from 44,100 to 22,050 Hz. Then, computed its STFT by sliding a Hamming window of 1411 samples with a 75% overlap to obtain the spectrogram.

3.2. Dictionary and E_0

Our implementation of Online Dictionary Learning for Sparse Coding (ODL) [12] is based on the SPAMS toolbox. Given N signals ($x_i \in \mathbb{R}_m$), ODL learns a dictionary D by solving the following joint optimization problem,

$$\begin{aligned} &\min_{D\geq 0,\alpha} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\left\|x_i - D\alpha_i\right\|_2^2 + \lambda\left\|\alpha_i\right\|_1\right) \\ &s.t.\ d_j^{\mathrm{T}}d_j \leq 1, \alpha_i \geq 0 \end{aligned} \tag{15}$$

where $\|\cdot\|_2$ denotes the Euclidean and λ is a regularization parameter. The input frames are extracted from the training set after short-time Fourier transform (STFT). Following [8], we define the dictionary size to be 100 atoms.

To get the reconstructed voice spectrogram from the annotation (E_0), we first transform the human-labeled vocal pitch contours into a time-frequency binary mask. The authors in [13] have proposed a harmonic mask similar to that of [14], which only passes integral multiples of the vocal fundamental frequencies [15] [16],

$$M(f,t)=\begin{cases}1, & \text{if } \left|f-nF_0(t)\right|<w/2, \exists n\in N^+\\ 0, & \text{otherwise.}\end{cases} \tag{16}$$

Here $F_0(t)$ is the vocal fundamental frequency at time t, n is the order of the harmonic, and w is the width of the mask. Then we simply define the vocal annotations as $E_0 = X \circ M$, where $\circ$ denotes the Hadamard product.

3.3. Evaluation

Separation performance is measured by BSS EVAL toolbox version 3.0[1]. We use source-to-interference ratio (SIR), source-to-artifacts ratio (SAR) and source-to-distortion ratio (SDR) provided in the commonly used BSS EVAL toolbox version 3.0. Denotes the singing voice $\hat{v}$, the original clean singing voice v, the source-to-distortion ratio (SDR) [17] is computed as follows,

$$\mathrm{SDR}(\hat{v},v)=10\log_{10}\left[\frac{\langle\hat{v},v\rangle^2}{\|\hat{v}^2\|\|v^2\|-\langle\hat{v},v\rangle^2}\right]. \tag{17}$$

Normalized SDR (NSDR) is the improvement of SDR from the original mixture x to the separated singing voice $\hat{v}$ [18] [19], and is commonly used to measure the separation performance for each mixture,

$$\mathrm{NSDR}(\hat{v},v,x)=\mathrm{SDR}(\hat{v},v)-\mathrm{SDR}(x,v). \tag{18}$$

For overall performance evaluation, the global NSDR (GNSDR) is calculated as,

$$\mathrm{GNSDR}=\frac{\sum_{i=1}^{N} w_i \mathrm{NSDR}(\hat{v}_i,v_i,x_i)}{\sum_{i=1}^{N} w_i}, \tag{19}$$

where N is the total number of the songs and w_i is the length of the i-th song. Higher values of SIR, SAR, SDR, GSIR, GSAR, GSDR and GNSDR represent better quality of the separation.

3.4. Parameter Selection

During parameter selection, we use the indicator of global normalized source-to-distortion ratio (GNSDR) as the evaluation index. The higher the val-

[1]http://bass-db.gforge.inria.fr/.

ue is, the better the separation quality is. In our algorithms, we set $\lambda_1 = \lambda_2 = 1/\sqrt{\max(m,n)}$ for each $X \in R^{m\times n}$ similar to [9], Here we only adjust γ.

Figure 1 presents the GNSDR for the separated singing voice and background music, using LSPDi. In the vocal part, we can see that, the GNSDR monotonically increases with γ first and then gradually decreases. When $\gamma = 5$, the LSRi achieves the overall highest GNSDR. In the accompaniment part, the values of GNSDR increase first, steady after $\gamma = 5$. Therefore, we set the parameter $\gamma = 5$.

3.5. Comparison Results

We compare three different Low-rank, Sparse algorithms on the iKala dataset,

- **RPCA** unsupervised method proposed by Huang *et al.* [3], use default parameter values $\lambda = \frac{1}{\sqrt{\max(m,n)}}$.
- **LSPD** Supervised method proposed by Yu *et al.* [4], use default parameter values $\lambda_1 = \lambda_2 = \frac{1}{\sqrt{\max(m,n)}}$.
- **LSRi** Proposed LSRi method with Low-Rank representation and the reconstructed voice spectrogram from the annotation, $\lambda_1 = \lambda_2 = \frac{1}{\sqrt{\max(m,n)}}$ and $\gamma = 5$.

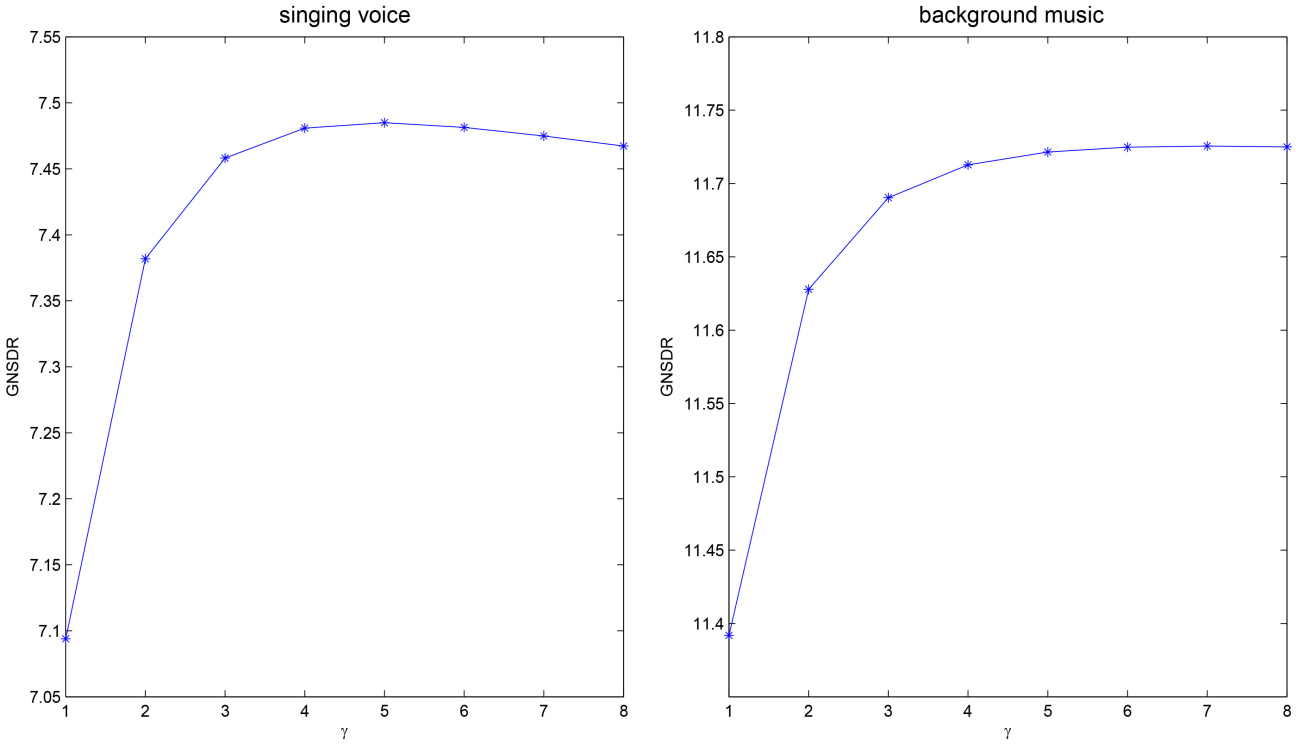

Figure 1. Separation performance measured by GNSDR for the singing voice (left) and background music (right), using our proposed method LSPDi.

Table 1. Separation quality for the singing voice and music for the iKala dataset of RPCA, LSPD and LSRi.

Method \ Index	Vocal			Music		
	GNSDR	GSIR	GSAR	GNSDR	GSIR	GSAR
RPCA	2.41	8.14	12.53	4.48	3.23	7.00
LSPD	1.45	11.47	7.19	4.95	2.47	11.73
LSRi	7.48	19.91	12.16	11.72	16.70	8.88

As shown in Table 1, whether the singing part or the accompaniment, our method has a higher value of global normalized source-to-distortion ratio (GNSDR), which suggests that LSRi algorithm performs well in the overall separation performance, and introduction of prior knowledge improve the separation performance. In the vocal part, our algorithm achieves higher GSIR than RPCA and LSPD, which shows that LSRi has better ability to remove the instrumental sounds than RPCA and LSPD. In the background music part, our algorithm achieves higher GSIR, which suggests that LSRi has better ability to remove the singing, a better performs in limiting artifacts during the separation process. But GSAR values did not improve significantly, this indicates that we need to improve on eliminating the interference of the algorithm.

4. Conclusion

In this paper, we have presented a time-frequency based source separation algorithm for music signals. LSRi considers both the vocal and instrumental spectrograms as sparse matrix and low-rank matrix, respectively. And the components that are not identified parts are specified as a residual term. Note that the dictionaries for the singing voice and background music are pre-learned from isolated singing voice and background music training data, respectively. Furthermore, LSRi incorporates vocal annotations information further, through which prior knowledge of the voice and background music is introduced to the source separation processing. Our approach has successfully exploited relevant useful information. Evaluations on the iKala dataset show the proposed methods better performance for both the separated singing voice and music accompaniment. In future studies, we can consider applying LSRi to the separation of complete songs.

References

[1] Li, Y. and Wang, D.L. (2007) Separation of Singing Voice from Music Accompaniment for Monaural Recordings. *IEEE Transactions on Audio, Speech and Language Processing*, **15**, 1475-1487. https://doi.org/10.1109/TASL.2006.889789

[2] Candes, E.J., Li, X., Ma, Y. and Wright, J. (2011) Robust Principal Component Analysis? *Journal of the ACM*, **58**, 1-37. https://doi.org/10.1145/1970392.1970395

[3] Huang, P.S., Chen, S.D., Smaragdis, P. and Johnson, M.H. (2012) Singing Voice Separation from Monaural Recordings Using Robust Principal Component Analysis.

2012 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Kyoto, 25-30 March 2012, 57-60. https://doi.org/10.1109/ICASSP.2012.6287816

[4] Yu, S., Zhang, H. and Duan, Z. (2017) Singing Voice Separation by Low-Rank and Sparse Spectrogram Decomposition with Pre-Learned Dictionaries. *Journal of the Audio Engineering Society*, **65**, 377-388. https://doi.org/10.17743/jaes.2017.0009

[5] Chan, T.S., Yeh, T.C., Fan, Z.C., Chen, H.W., Su, L., Yang, Y.H. and Jang, R. (2015) Vocal Activity Informed Singing Voice Separation with the iKala Dataset. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brisbane, 19-24 April 2015, 718-722. https://doi.org/10.1109/ICASSP.2015.7178063

[6] Lehner, B., Widmer, G. and Sonnleitner, R. (2014) On the Reduction of False Positives in Singing Voice Detection. 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Florence, 4-9 May 2014, 7480-7484. https://doi.org/10.1109/ICASSP.2014.6855054

[7] Yoshii, K., Fujihara, H., Nakano, T. and Goto, M. (2014) Cultivating Vocal Activity Detection for Music Audio Signals in a Circulation Type Crowd Sourcing Ecosystem. 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Florence, 4-9 May 2014, 624-628. https://doi.org/10.1109/ICASSP.2014.6853671

[8] Chan, T.S. and Yang, Y.H. (2017) Informed Group-Sparse Representation for Singing Voice Separation. *IEEE Signal Processing Letters*, **24**, 156-160.

[9] Chan, T.S., Yeh, T.C., Fan, Z.C., Chen, H.W., Sui, L., Yang, Y.H. and Jang, R. (2015) Vocal Activity Informed Singing Voice Separation with the iKala Dataset. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brisbane, 19-24 April 2015, 718-722. https://doi.org/10.1109/ICASSP.2015.7178063

[10] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016

[11] Ma, S. (2016) Alternating Proximal Gradient Method for Convex Minimization. *Journal of Scientific Computing*, **68**, 546-572. https://doi.org/10.1007/s10915-015-0150-0

[12] Mairal, J., Bach, F., Ponce, J. and Sapiro, G. (2009) Online Dictionary Learning for Sparse Coding. *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, 14-18 June 2009, 689-696. https://doi.org/10.1145/1553374.1553463

[13] Ikemiya, Y., Yoshii, K. and Itoyama, K. (2015) Singing Voice Analysis and Editing Based on Mutually Dependent F0 Estimation and Source Separation. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brisbane, 19-24 April 2015, 574-578. https://doi.org/10.1109/ICASSP.2015.7178034

[14] Virtanen, T., Mesaros, A. and Ryynanen, M. (2008) Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music. *ITRW on Statistical and Perceptual Audio Processing*, Brisbane, 21 September 2008, 17-22.

[15] Durrieu, J.L., David, B. and Richard, G. (2011) A Musically Motivated Midlevel Representation for Pitch Estimation and Musical Audio Source Separation. *IEEE Journal of Selected Topics in Signal Processing*, **5**, 1180-1191.

https://doi.org/10.1109/JSTSP.2011.2158801

[16] Ryynanen, M., Virtanen, T., Paulus, J. and Klapuri, A. (2008) Accompaniment Separation and Karaoke Application Based on Automatic Melody Transcription. 2008 *IEEE International Conference on Multimedia and Expo*, 23 June-26 April 2008, Hannover, 1417-1420.

[17] Gribonval, R., Benaroya, L., Vincent, E. and Fvotte, C. (2003) Proposals for Performance Measurement in Source Separation. 4*th International Symposium on Independent Component Analysis and Blind Signal Separation* (*ICA*2003), Nara, April 2003, 763-768.

[18] Ozerov, A., Philippe, P., Gribonval, R. and Bimbot, F. (2005) One Microphone Singing Voice Separation Using Source-Adapted Models. 2005 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, 16 October 2005, 90-93. https://doi.org/10.1109/ASPAA.2005.1540176

[19] Ozerov, A., Philippe, P., Bimbot, F. and Gribonval, R. (2007) Adaptation of Bayesian Models for Single Channel Source Separation and Its Application to Voice/Music Separation in Popular Songs. *IEEE Transactions on Audio, Speech and Language*, **15**, 1564-1578. https://doi.org/10.1109/TASL.2007.899291