RESEARCH ARTICLE OPEN ACCESS

Separation of Singing Voice from Arabic Music Background Using Robust Principal Component Analysis

M. E. ElAlami [1], S.M.K.tobar [2], S.M.Khater [3], Eman.A.Esmaeil [4]
Computer Science Department [1], [3], [4], Musical Education Department [2]
Faculty of Specific Education, Mansoura University
Egypt

ABSTRACT

Separation of voice singing has gained significant interest in many applications in the world. The aim to separate music and singing voice approach is to isolate music signal from the voice of the sing, which consider an important music information retrieval (MIR) technology. To process wide music libraries, tools can extract important details about musical pieces from music signal are needed .This paper discusses a proposed method for using Robust Principal Component Analysis technique to separate voice of singing from Accompanying arabic music. Result show that RPCA method achieves important success for Arabic Music Background separation from Voice of sing.

Keywords :— Arabic Music background , Robust Principal Component Analysis, Time frequency masking, music information retrieval

I. INTRODUCTION

A useful knowledge for song gives by a singing voice, because it embodies the singer, the lyrics and the song's emotions. This information had been used by several systems such as automated text recognition, singer detection, retrieval of musical information, karaoke applications, classification of musical genres, melodic extraction and separation of polyphony music. [1-2]

However, the existing separation techniques still far away from ability of human hearing [3].

There are major challenges to current problems in voice-singing, like: [4]

- The scene created by music can generally be included multisource, where various sources of sound from different types of instruments are temporarily involved, some just slightly.
- The source of music can have multiple instrumental levels, can be played at varying loudness and pitches, and even the spatial location of a sound source can vary over time.
- The singing voice has a much different frequency of pitch for males and females who can overlap with the background frequency pattern at a certain point
- Previous studies can be divided into two classes on the singing voice separation frameworks: [5]
- The supervised methods are the first class, which initially map a space of feature to detect voice of sing , and recently apply techniques such as pitch interference [6 -7].
- The Unsupervised methods are the second class, who request no previously training features, like filter model [8].

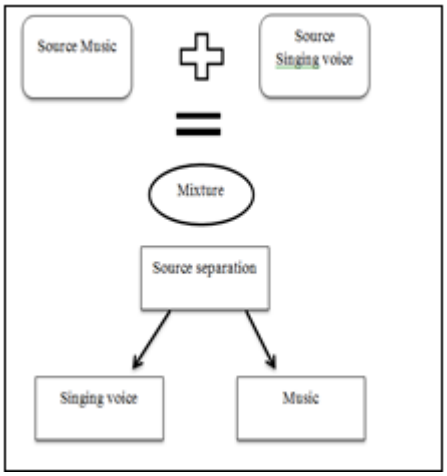

Figure 1: Overview Diagram of Singing Voice Separation process

The singing voice is usually the main melodic source in popular music; therefore, the most difficult task is to detect and extract the voice of sing from music signal .The extracted melody is important in several respects, including classifying various music genres. [9]

This paper is ordered as following : Section2 addresses related work, Section 3 discuss the proposed system. Section 4 discusses our experimental result, Section 5 discusses the results and the summary is presented in Section 6.

II. RELATED WORK

The significance of programmed music has been developing without interruption until two years. various papers suggested systems interested in singing voice separating from musical background and using machine learning algorithms to create models for music classses .Music source separation is the job of decompose music into its constitutive components, and it has a long history of scientific activity as it is known to be a very challenging problem, so, many techniques have been suggested with the goal of overcoming the hardness in separation task.

British et al.[10] developed a novel process for extracting vocal track from a mixture of material. The musical mix consists of a voice singing and a rhythm track that may consist of various instruments. Then, the approximate parameters are used to synthesize vocal sound, without any back sound intervent .

The result shows that this system was among the best algorithms in the MIREX2016 source separation competition.

Feng and Masato. [11] proposed an unsupervised voice separation algorithm dependent gammatone auditory filter bank expansion of robust component analysis with rank -1 restriction .

The result shows that better separation efficiency on MIR-1 K dataset can be achieved with the proposed algorithm.

Zafer and Bryan[12], provided a clear method for separating music and voice to eliminating the repeated musical structure,. The system includes:

(1) Repeating structure time is detected.

(2) Spectrogram is segmented and the segments are averaged to construct a pattern of repeated segments.

(3) Model is matched with each time-frequency in a section and the mixture is split using binary time frequency masking.

The evaluation result shows that this approach can boost the efficiency of an established music / sound separation by using a data of 1,000 song clips.

Bryan et al. [13] introduced an approach monitoring the repeated calculated local context pattern estimates and structure time and Separation is achieved by masking time frequency, due to the difference between the current observation and the predicted background trend.

The results of an evaluation on a dataset of 14 completed tracks show that this approach consider a competitive method of music separation and being computationally efficient.

Andreas et al. [14] proposed a novel application of the U-Net architecture for the task of separation, despite its proven ability to reproduce the precise, low-level detail required for high-quality audio reproduction.

The result shows that the proposed algorithm achieves state of the art efficiency by both quantitative evaluation and subjective evaluation.

Gerard et al.[15] provided a qualified convolutionary DNN to to get estimation of the ideal binary mask for the separation of vocal sounds from mixtures of music of the real world.

The result shows that this method can be useful for extracting vocal automatically from mixtures music for applications of 'karaoke' type.

DeLiang et al. [16] introduced an algorithm for pattern estimation to discover the pitch ranges of voice singing within each time frame. The result shows that this algorithm outperform preceding systems for pitch extraction and voice separation singing.

Abouzid and Chakkor[17] proposed a novel approach for signal separation algorithms based on Gaussianity and Sparsity where independent analysis of the components would be used. The Sparsity as a pre-processing step, then the Gaussianity - based source separation block was used as a final stage for estimating the original sources. The result shows that FPICA algorithm acheive best evaluation.

Ying and Guizhong [18] introduced a two-phase system. The first stage uses the non-negative partial co-factorization matrix (NMPCF) In the second stage, the pitches of voice of sing are first determined on the basis of the separate voice obtained from the first stage.

The Experimental results show that singing voice refinement would further improve Delta SNR from the standpoint of source separation compared to voice sing separation by NMPCF, while voice sing separated by NMPCF is more appropriate for singer recognition than refined singing voice.

III. THE PROPOSED SYSTEM

This paper discusses arabic Music Background separation method from voice of singing .One of the most effective methods used for process of separation is robust principal component analysis (RPCA). In our proposed system the vocal signal has sparse distribution and non-stationary feature, While the sound of arabic music background is really continuous and Repeatable. The voice separation flowchart is shown in Figure 2

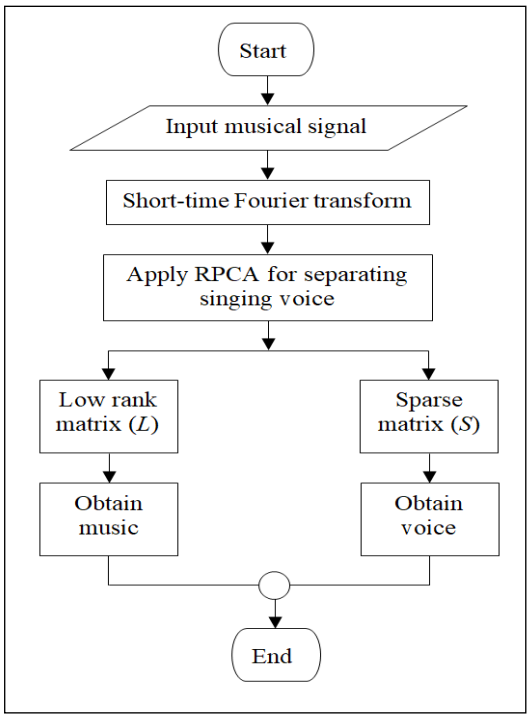

Fig 2. Flowchart for proposed separation system

The main steps of system of separation are illustrated as follows;

3. 1. Short-Time Fourier Transform (STFT)

The short-time Fourier transform (STFT) is extracted from the signal input where the signal in the time frequency domain is defined. (STFT) which is considered the spectrogram, is extracted from a normal Fourier transform by multiplying the time signal $s[g]$ by a sliding time window necessary. $w[J]$. The sliding window position adds a time dimension and one gets a time-varying overview of the frequency.

STFT is described as: [19]

$$S[A,E]=\sum_{n} s[g]w[J-f]e^{-j2\pi mk/C} \qquad (1)$$

As $S[A,E]$ is the SFTF of the discrete signal $s[n]$, $w[n]$ is a short-time windowing function of size n, centered at time location f and C Is the discrete frequency number. Since the Fourier transformation is a complicated function, the density of the power spectrum $P_s[A,E]$ is described as [20] :

$$P_s[A,E]=\frac{1}{C}|S[A,E]|^2 \qquad (2)$$

for sampling frequency f_s, every windowed frame is introduction by C-points power spectrum covering the range of frequency $[-f_s/2, f_s/2[$. The spectrum of power do not use as a feature vector directly since it include data of large size (C/2 components). By using of overlapping hamming window with C=1024 samples at a sampling rate of 16 KHz, STFT of the input audio signal is calculated in the process of separation.

The signal under analysis is divided to a number of small tracks in the spectrogram , where it is presumed that each sub-tracks is steady. each sub-tracks is multiplied by a suitable range to leak the effect finite data , Afterwards the FFT technique had applied to every sub-track. Because of each sub-set contain no unexpected changes, the spectrogram can provide an perfect understanding of how the signal's spectral composition has changed over all the entire time span. and yet, there are signals in nature which spectral content is so quickly evolving that it is hardly to find a suitable short-term window, since no term interval over which the signal is stationary can exist. To accommodate these shifts in time properly, the length of the time span should be kept as short as probable. However, this will increase the resolution of frequencies on the plane of time-frequency. And there will be a comparison between time resolutions and Resolutions of frequency

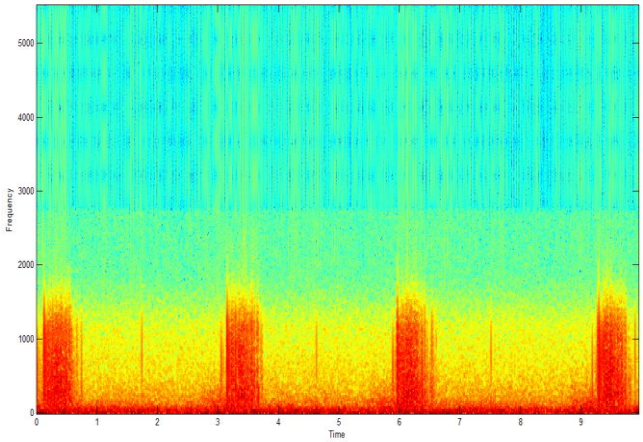

Figure 3: Spectrogram Process

3. 2. Applying RPCA for separation of Singing voice.

The RPCA is implemented as an optimization technique via the Augmented Langrange Multiplier (ALM), that decipheres the RPCA computational problem. The music accompaniment can help musical instruments to repeat the same sounds in the same music all time , So its value spectrogram can beknown as a low-rank matrix structure. In addition to, the part of voice sing significantly varies andand has a sparse spectrogram of classification field because of its harmonic form, leading to a spectrum with a sparse matrix form. Using

the RPCA techniques, an input matrix is decompose to a sparse matrix and a low rank matrix (melody accompaniment).The method of Principal Component Approach, proposes to solve the following issue of convex optimisation [22]

$$\text{Minimize } \|S1\|* + \lambda\|S2\|1 \qquad (3)$$

$$\text{Subject to } S1 + S2 = M$$

As $M \in Rn1\times n2$, $S \in Rn1\times n2$, $L \in Rn1\times n2$, $\|\cdot\|*$ and $\|\bullet\|1$ Denoting the nuclear norm (single value sumition) and the L1 norm (Summation of absolute matrix entry values), Accordingly. $\lambda > 0$ is a parameter of trade-off between the range L and the sparsity S.

By using a value of $\lambda = 1/\sqrt{\max(n1, n2)}$ Is a strong rule which can then be produce the best possible result. λ k = k/max (n1, n2).

Numerous Traffic variations between S and L rank were tested with different k values.

Because of Instrumental music can repeat the Similar sound when they are played , Music usually has an underlying repetitive musical form, music is considered a signal of lower rank. Onto the reverse, singing voices have more range, but in time and frequency domains they are quite scarce. Speech singing is an essential part of a sparse matrix. by using RPCA, The low-rank matrix L includes music background , and the sparse matrix S includes vocal signals which can be generated by two matrices output S and L.

3. 3. TIME-FREQUENCY MASKING

Masking can be applied to ALM's results of separation which include low rank matrix and sparse matrix by using binary time frequency masking for better separation output for matching the song's rhythmic structure with masking for enhanced separation efficiency, music signals need to be precisely separated as voice of sing often lines the music background during beat instances. Masking time frequency is measured by this Equation:

$$M_b(y1, y2) = \begin{cases} 1 & |S(y1, y2)| > gain * |L(y1, y2)| \\ 0 & otherwise \end{cases} \qquad (4)$$

$$for\ all\ \ y1 = 1...n_1\ \ and\ \ y2 = 1...n_2$$

When the time-frequency mask is calculated, its applied in STFT matrix , to result two matrix

matrix $E_{singing}$ and E_{music}, as in the following equation.

$$E_{singing}(y1, y2) = M_b(y1, y2)\ M(y1, y2)$$
$$E_{music}(y1, y2) = (1 - M_b(y1, y2)\, M(y1, y2) \qquad (5)$$

As y1 = 1. . . n1 and y2 = 1. . . n2.

After applied time frequency masking, It is applied onto the main audio signal to present separation matrix as a voice sing and musical background.

3. 4. Inverse Short Time Fourier Transform

At last, Short Time Fourier Transform (ISTFT) can be used to generate the expected results waveform assisted by the test evaluation [24].Figure 4 shows the steps of RPCA and The RPCA for separation algorithm is shown in figure 5.

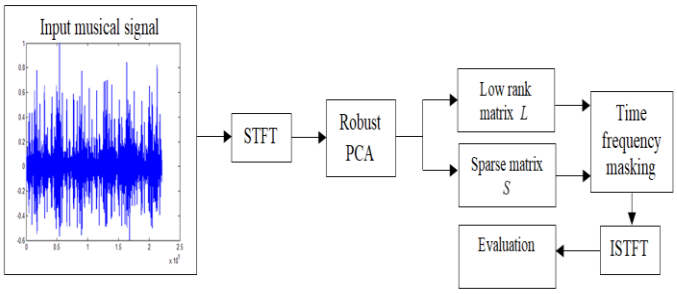

Fig 4. Block diagram of singing voice separation using RPCA

Algorithm RPCA for singing voice separa (23)

Input: Mixture signal $M \in \mathbb{R}_{m\times n}$.
1: **Initialize**: $\rho > 1, \mu_0 > 0, k = 0, L_0 = S_0 = 0$.
2: While not convergence,
3: **do** :
4: $L_{k+1} = P_{1,\mu_k^{-1}}(M - S_k + \mu_k^{-1} J_k)$.
5: $S_{k+1} = Q_{\lambda\mu_k^{-1}}(M - L_{k+1} + \mu_k^{-1} J_k)$.
6: $J_{k+1} = J_k + \mu_k(M - L_{k+1} - S_{k+1})$.
7: $\mu_{k+1} = \rho * \mu_k$.
8: $k = k + 1$.
9: **end while**.
Output: $L_{m\times n}$, $S_{m\times n}$.

Figure 5: The Algorithm of RPCA for separation process

IV. EXPERIMENTAL RESULT

RPCA algorithm applied to database of Arabic songs, which consisting of male and female singers with a sample rate of 16Khz and audio clip length of 60-90 seconds. Robust Principal Component Analysis suggests That the background music lies in the low-rank subspace, when the voice is singing

is comparatively sparse because of its greater variability in the song itself. The proposed approach is an algorithm for matrix factorisation to solve sparse matrix and low-rank matrix .The data were extracted from Arabic music songs, having more than two background musical instruments, used to evaluate outputs of separation music system .(STFT) and inverse STFT (ISTFT) are used to calculate features. A hop size of 256 samples and an FFT size of 1024 a window size of 1024 samples is used.

The audio files separated are compared with those files for the valuation of the results. For testing the efficacy of the proposed approach in words of the source to artifact ratio (SAR) and source to distortion ratio (SDR) by using of BSS-EVAL 3.0 metrics, which is a MATLAB toolbox , to calculate the efficiency of (blind) source separation techniques within an estimate framework. BSS-EVAL 3.0 metrics is determined as [25]:

$$S(t) = S_{target}(t) + S_{interf}(t) + S_{artif}(t) \qquad (6)$$

As S Interf (t) is deformation permissible of the sources accounting for the intervention of the undesired sources; S artif (t) is an artifact term which may mention to the artifact of the technique of separating ; and S target (t) is the acceptable deformation of the Aim tone.

The SDR and SAR formulae are described as [26]

$$SDR = 10\log 10 \frac{\sum_t S_{target}(t)^2}{\sum_t \{e_{interf}(t)+e_{artif}(t)\}^2} \qquad (7)$$

The SAR is defined as:

$$SAR = 10\log 10 \frac{\sum_t \{s_{target}(t)+e_{interif}(t)\}^2}{\sum_t e_{artif}(t)^2} \qquad (8)$$

The Normalized SDR (NSDR) is determined as:

$$NSDR(\bar{v}, v, x) = SDR(\bar{v}, v) - SDR(x, v) \qquad (9)$$

Where v is the original voice of clear singing, $\bar{v}$ is the resynthesized voice of singing and x is the mixture. NSDR is intended to estimate the change of separate sound of the singing $\bar{v}$ and pre-processed mixture x.

In addition to evaluating the terms of Source to Objects Ratio (SAR) , Source to Distortion Ratio (SDR) and Source to Interference Ratio (SIR) by BSS-EVAL metrics GNSDR is used for testing our system. The GNSDR is determined by getting the NSDRs average over all mixtures of each collection, weighted by its duration. Described of GNSDR showed as [27]

$$GNSDR(\bar{v}, v, x) = \frac{\sum_{n=1}^{N} W_n \, NSDR(V_n, V_n, X_n)}{\sum_{n=1}^{N} W_n} \qquad (10)$$

Where N is the total track number, w_n is the duration of the nth track and n is a song index.

Higher values for SDR, SIR and NSDR indicate that our approach provides improved separation efficiency. SDR stands for the distinct target signal level, SIR represents a divided level between the targets, NSDR can be used to estimate overall separation efficiency.

Result show that our proposed system achieves a high separation rate arrive to 98%.

Figures blew showed experimental RPCA result for first 40 second of 3-salamat arabic song .

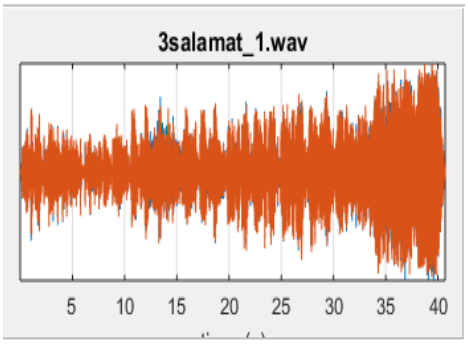

Fig.6 Original signal for first 40 second

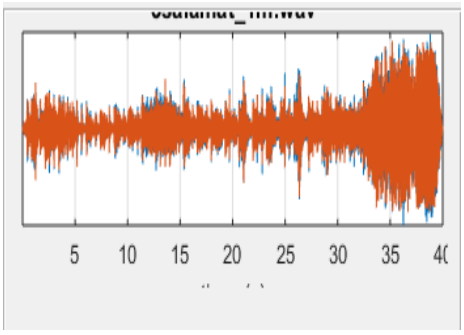

Fig.7 Music Signal for first 40 second of 3-salamat Song after applying RPCA

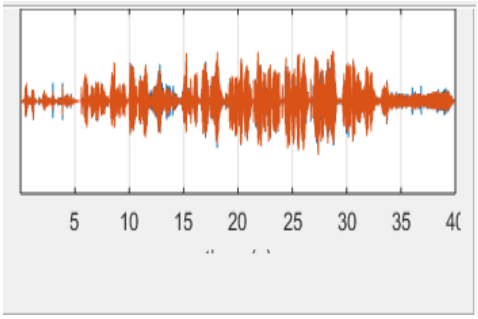

Fig.8 Speech signal for first 40 second of 3-salamat Song after applying RPC

TABLE 1

Evaluation table for 10 song files

Song number	SDR	SIR	SAR	GNSDR
1	2.7360	5.0676	32.8097	2.6610
2	3.2495	7.0665	32.1440	3.2282
3	3.0895	6.7545	26.4082	3.2028
4	2.3568	4.9963	24.4931	2.4452
5	2.5535	8.8622	28.1980	2.5818

V. CONCLUSION

In this paper, a supervised approach to the task of vocal separation singing is discussed. This approach has proven to be suitable for many musical formats as our experience on the Arabic song database improves. In the future work, it will be suggested to combine melody extraction with the music mix signal to improve the class results and focus on the structural and temporal features of frequency because it is necessary to completely isolate the singing sound from the mixing music signal.

REFERENCES

[1] Mesaros , A. and Virtanen, T. (2010). Automatic Recognition of Lyrics in Singing. EURASIP Journal on Audio, Speech, and Music Processing, 2010, pp.1-11.

[2] A. J. R. Simpson, G. Roma, and M. D. Plumbley, Deep karaoke: n extracting vocals from musical mixtures using a convolutional deep neural network, in Proc. LVA/ICA, pp.

[3] Asim, M. and Ahmed, Z. (2017). Automatic Music Genres Classification using Machine Learning. International Journal of Advanced Computer Science and Applications, 8(8).

[4] Burute, H. and Mane, P. (2015). Separation of Singing Voice from Music Background. International Journal of Computer Applications, 129(4), pp.22-26.

[5] Z. Rafii and B. Pardo, A simple music/voice separation method based on the extraction of the repeating musical structure, in ICASSP, May 2011, pp. 221– 224.

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

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

[8] Z. Rafii and B. Pardo, A simple music/voice separation method based on the extraction of the repeating musical structure, in *ICASSP*, May 2011, pp. 221– 224.

[9] Kum , S. and Nam, J. (2019). Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks. Applied Sciences, 9(7), p.1324.

[10] Chandna, P., Blaauw, M., Bonada, J. and Gomez, E. (2019). A Vocoder Based Method for Singing Voice Extraction. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[11] Li, F. and Akagi, M. (2018). Unsupervised Singing Voice Separation Using Gammatone Auditory Filterbank and Constraint Robust Principal Component Analysis. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

[12] Rafii, Z. and Pardo, B. (2011). A simple music/voice separation method based on the extraction of the repeating musical structure. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[13] Liutkus, A., Rafii, Z., Badeau, R., Pardo, B. and Richard, G. (2012). Adaptive filtering for music/voice separation exploiting the repeating musical structure. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[14] Jansson , A., Humphrey, E. (2017). singing voice separation with deep u-net convolutional networks. Proceedings of the 18th ISMIR Conference, Suzhou, China, October 23-27, 2017.

[15] Simpson, A., Roma, G. and Plumbley, M. (2015). Deep Karaoke: Extracting Vocals from Musical Mixtures Using

a Convolutional Deep Neural Network. Latent Variable Analysis and Signal Separation, pp.429-436.

[16] Hsu, C., Wang, D., Jang, J. and Hu, K. (2012). A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment. IEEE Transactions on Audio, Speech, and Language Processing, 20(5), pp.1482-1491.

[17] Abouzid, H. and Chakkor, O. (2017). Blind Source Separation Approach for Audio Signals based on Support on Audio, Speech, and Language Processing, 20(5), pp.1482-1491.

[18] Hu, Y. and Liu, G. (2015). Separation of Singing Voice Using Nonnegative Matrix Partial Co-Factorization for Singer Identification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(4), pp.643-653.

[19] Bahoura, M., 2019. Efficient FPGA-Based Architecture of the Overlap-Add Method for Short-Time Fourier Analysis/Synthesis. Electronics, 8(12), p.1533.

[20] Bahoura, M. and Ezzaidi, H., 2012. FPGA implementation of a feature extraction technique based on

Fourier transform. 2012 24th International Conference on Microelectronics (ICM).

[21] Debbal, S. and Bereksi-Reguig, F., 2007. Time-frequency analysis of the first and the second heartbeat sounds. Applied Mathematics and Computation, 184(2), pp.1041-1052.

[22] Huang, P., Chen, S., Smaragdis, P. and Hasegawa-Johnson, M. (2012). Singing-voice separation from monaural recordings using robust principal component analysis. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[23] Umap, P. and Chaudhari, K. (2015). Singing Voice separation from Polyphonic Music Accompanient using Compositional Model. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 04(02), pp.541-546.

[24] F. Li and M. Akagi, "Unsupervised singing voice separation based on robust principal component analysis exploiting rank-1 constraint," in Proc. EUSIPCO, 2018, pp. 1929-1933.

[25] E. Vincent, R. Gribonval, and C. F´evotte, "Performance measurement in blind audio source separation," IEEE Trans. ALSP, vol. 14, no. 4, pp. 1462-1469, 2006.

[26] Kornycky, J., Gunel, B. and Kondoz, A. (2008). Comparison of Subjective and Objective Evaluation Methods for Audio Source Separation

[27] Ikemiya, Y., Itoyama, K. and Yoshii, K. (2016). Singing Voice Separation and Vocal F0 Estimation Based on Mutual Combination of Robust Principal Component Analysis and

Subharmonic Summation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11), pp.2084-2095.