International Journal of Innovative
Computing, Information and Control
Volume **14**, Number **6**, December **2018**

 ISSN 1349-4198
pp. **2325–2333**

RENICA BASED MUSIC SOURCE SEPARATION FOR AUTOMATIC MUSIC EMOTION CLASSIFICATION

Nurlaila Rosli, Nordiana Rajaee and David Bong

Department of Electrical and Electronic Engineering
Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia
laila8805@gmail.com; { rnordiana; bbldavid }@unimas.my

Received February 2018; revised June 2018

Abstract. *In music source separation, we deal with the problem of precisely separating the singing voice and instrumental accompaniment estimation in music mixtures. This problem is addressed through the use of signal processing algorithm called RENICA which refers to the combination of source separation methods namely, REpeating Pattern Extraction Technique (REPET), Nonnegative Matrix Factorization (NMF) and Independent Component Analysis (ICA). The separated estimation is later used for parameters modelling in automatic Music Emotion Classification (MEC). This paper aims to improve singing voice and instrumental accompaniment separation in 120-180 sec music signal length by merging three music source separation algorithms. From the experimental results obtained, this new combination of algorithms not only succeeds in separating better estimation but enhances the accuracy for emotion based music classification up to 97% for angry, peaceful, happy and sad emotion categories.*
Keywords: Music source separation, REPET, Nonnegative matrix factorization, Independent component analysis, Music emotion classification

1. **Introduction.** Music is a language of emotions [1]. Recently, emotion based music classification is regarded as one of the important researches in various fields including music information retrieval, psychology, active music listening and affective computing [2-4]. Musical content can be extracted and analyzed to model the emotion parameters in music for the classification. Many researchers also have been using musical timbre features in MEC for better accuracy. However, the incorporation of timbre features in both singing voice and instrumental accompaniment for Music Emotion Classification (MEC) is still in its infancy [5].

The exploitation of timbre features parameters in singing voice for automatic MEC became less affective if mixed with the instrumental accompaniment. Thus, the separation of both singing voice and instrumental accompaniment in MEC is crucial and failure to estimate the overlapping partials and separating musical component in the mixture may result in inaccurate classification.

Most of the source separation algorithms have been widely applied in solving numbers of problem in speech processing, speech denoising and 'Cocktail Party Problem'. However, due to the rapid growth in digital technology and music information retrieval, the use of source separation techniques has been expanded to cater problems which is related to music. For the past few years, we have witnessed various algorithms, manipulated to separate music mixture into estimated sources, including the latest REpeating Pattern

DOI: 10.24507/ijicic.14.06.2325

Extraction Technique (REPET), Robust Principal Component Analysis (RPCA), Nonnegative Matrix Factorization (NMF), Independent Component Analysis (ICA) and Blind Source Separation (BSS).

In [6], Rafii and Pardo proposed REpeating Pattern Extraction Technique (REPET) as source separation algorithms that estimate singing voice by identifying and comparing the repeating segments before the extraction of the repeating patterns. REPET works similarly to a drum sound recognizer where the time-frequency segments are repeatedly updated according to the drum patterns in spectrogram [7]. Based on work done in [8], Nonnegative Matrix Factorization (NMF) is an established source separation method that has been used to separate signals into their instrumental and singing voice parts by modeling the spectral components into time-frequency segments.

Independent Component Analysis (ICA) is a matrix factorization method utilized to decompose the observation vector into statistically independent variables and it performs well for the convoluted mixture [9]. Blind Source Separation (BSS) [10] is an effective algorithm in ICA and requires no spatial or spectral information about the music mixture as it is assumed that the source signals have no correlations to each other. In BSS, the signals are separated into different sets of signals where the regularity between the signals is minimized and regularity of each signal is maximized [11].

In this study, we articulate our main contribution in music source separation method namely RENICA (REPET+NMF+ICA), where we estimate and separate two main sources in music which are Singing Voice (SV) and Instrumental Accompaniment (IA). Timbre factors are extracted and as expected, the threshold values of timbre for every SV and IA sources are found to be at variance. The comparison is done by training Multilayer Neural Networks (MLNN) with and without the exploitation of RENICA. The proposed method is programmed to accurately classify emotion in selected music.

2. Proposed Framework.

We highlight our proposed method as shown in Figure 1. RENICA is built from the combination of three music source separation algorithms which are REpeating Pattern Extraction Techniques (REPET), Nonnegative Matrix Factorization (NMF) and Independent Component Analysis (ICA). These algorithms are proven to be excellent in generating better estimations for much longer signal [12]. A thousand song clips with different emotion attributes were used in this experiment.

2.1. Music source separation using RENICA.

RENICA as shown in Figure 2 is a set of algorithm which consists of four important processes called IEDE referring to segment identification, patterns estimation, segments decomposition and sources extraction. Throughout this source separation process, the singing voice and instrumental accompaniment segments need to be identified. After identification, various techniques such

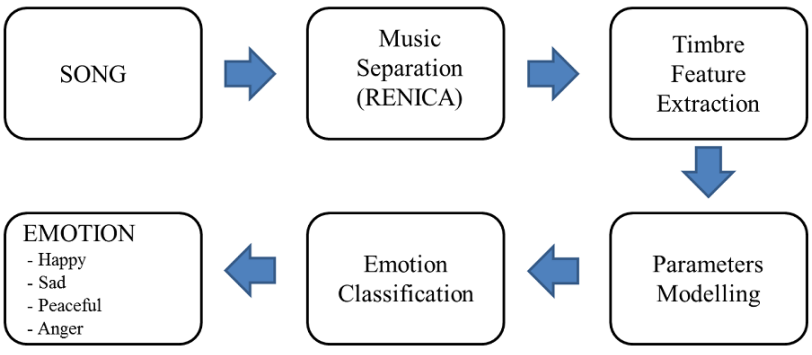

Figure 1. Proposed framework with RENICA

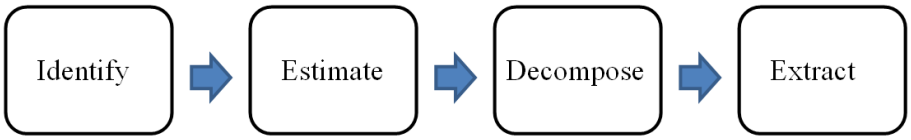

FIGURE 2. RENICA processes

as spectrogram factorization, accompaniment model learning and pitch based inference techniques are utilized to separate these segments.

Segment Identification: We started our separation process by identifying repeating segments in 120-180 sec (44.1 Hz) music mixtures. In RENICA, we separate the repeating segments and non-repeating segments by using autocorrelation to measure similarity between segments. The repeating beat spectrum b is calculated using Equation (1). The magnitude spectrogram V is derived after Short Time Fourier Transform (STFT) is evaluated using half overlapping Hamming windows of N samples. The element-wise square V^2 of each row is computed to obtain the beat matrix, matrix B. Further details of this calculation can be seen in [6,7].

$$\begin{aligned} &B(i,j) = \frac{1}{m-j+1} \sum_{k-1}^{m-j+1} V(i,k)^2 V(i,k+j-1)^2 \\ &\text{if } b(j) = \frac{1}{n} \sum_{i=1}^{n} B(i,j) \\ &\text{then } b(j) = \frac{b(j)}{b(1)} \\ &\text{for } i = 1,\ldots,n \text{ (frequency), where } n = \tfrac{N}{2} + 1 \\ &\text{for } j = 1,\ldots,m \text{ (lag), where } m =\neq \text{ time frames} \end{aligned} \tag{1}$$

Estimate: We estimate the repeating pattern by following two main conditions. 1) Constantly repeating segments is regarded as instrumental accompaniment and 2) non repeating segment is regarded as singing voice. The calculation for segment estimation is derived in the following Equation (2) using median model as used in [13]. We do this step to calculate the repeating (IA) and nonrepeating segments (SV) so we can initiate the learning rule and after every iteration we try to keep the vector orthogonal to the previously estimated sources.

$$\begin{aligned} &S(i,j) = \text{median}_{k=1,\ldots,r}\{V(i,l+(k-1)p)\} \\ &\text{for } i = 1,\ldots,n \text{ (frequency)} \\ &\text{and } l = 1,\ldots,p \text{ (time)} \\ &\text{where } p = \text{period length and } r =\neq \text{ segments} \end{aligned} \tag{2}$$

Decompose: We then decompose the repeating segments into two main sources: Singing Voice (SV) and Instrumental Accompaniment (IA). NMF and ICA-BSS approach is used to decompose both SV and IA segments. We assume that the magnitude spectrogram of mixed signals is a matrix of a non-negative number as proposed in [8].

The independent BSS decomposes the mixed signal with repeating patterns, $X(t) = (X_1(t),\ldots,X_m(t))^T$ by finding and using coefficients or activation vector, $H = [H_{ij}] \in R^{n\times m}$ to get the original SV or IA estimates, $W(t) = (W_1(t),\ldots,W_n(t))^T$. We use Equation (3) to obtain an update for W. If the change for $D(X||WH)$ is small from the last iteration, then declare convergence and return the value of updated W and H. The Kullback-Leibler – NMF algorithm is summarized in Equation (3).

$$D(X||WH) = D\left(X^T||H^T W^T\right) \tag{3}$$

Algorithm KL – NMF
Initialize **W**, **H**
Repeat
$H \leftarrow H * \dfrac{W^T \frac{x}{WH}}{W^T 1}$
$W \leftarrow W * \dfrac{\frac{x}{WH} H^T}{1 H^T}$
Until convergence
Return **W**, **H**

Extract: We extract singing voice and instrumental accompaniment segments using Equation (4), where s is a source, W_s is a basis vector, and H_s is an activation vector. The estimates of singing voice and instrumental accompaniment are obtained from $W_{SV}H_{SV}$ and $W_{IA}H_{IA}$.

$$\left|\hat{X_s}\right| = W_s H_s = \sum_{i \in s} \left(w_i h_i^T\right) \tag{4}$$

2.2. **Timbre feature extraction.** Timbre is the tone color that is determined by the harmonic and sound quality in music sources [14]. Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT) are applied to the windowed signals to computing the following factors.

MFCC: Mel-frequency cepstrum coefficient refers to the modeling of human auditory perception system as explained in [15]. Fourier transform of a signal is identified before the power spectrum is obtained through triangular overlapping windows. The log of powers at each of the mel-frequencies is taken and discrete cosine transforms of the mel log powers are extracted. The amplitudes of the resulting spectrums are referred to as MFCCs [16].

Spectral Centroid: Average frequency weighted by amplitudes is shown in Equation (5). Spectral centroid is utilized to measure the center of mass for a particular spectrum.

$$\text{Spectral centroid} = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]} \tag{5}$$

Spectral Roll off: The roll-off frequency is used to distinguish harmonic and noisy sounds. In this context, the frequency k is below 85% of the x magnitude distribution as shown in Equation (6).

$$\sum_{n=0}^{k} x(n) = 0.85 \sum_{n=0}^{N-1} x(n) \tag{6}$$

Low Energy: In audio signal processing and speech classifications, the energy curve is used as an assessment of the temporal distribution of energy in order to determine whether it remains constant throughout the signal. Spectral energy density is derived in Equation (7).

$$E_s(f) = |X(f)|^2 \tag{7}$$

Harmony Entropy: An estimation of entropy is given in Equation (8).

$$H(x) = -\int_x f(x) \log f(x) dx \tag{8}$$

Irregularity: It refers to irregularity of the spectrums, given as in Equation (9).

$$I = \sum_{k=2}^{N-1} \left| a_k - \frac{a_{k-1} + a_k + a_{k+1}}{3} \right| \tag{9}$$

Zero Cross: Zero cross indicates the amount of noise based on the number of times the signal crosses the zero line. Zero cross rate is given in Equation (10).

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(s[n-1])| \tag{10}$$

2.3. **Machine learning model.** Multilayer Neural Network (MLNN) with a supervised feed-forward backpropagation network is used as classifier in this work. MLNN consists of seven input neurons, 20 hidden neurons and single output neurons to train the neural network classifier. Summarization of backpropagation algorithm is as follows. Refer [17] for details.

Algorithm neural network backpropagation
  Initialize the weights and biases
    Feed the training sample
    Propagate the inputs forward
    Compute the net input and output of each unit in the hidden and output layers
    Backpropagate the error
    Update the weights and biases to
    Reflect the propagated errors
  Repeat and apply terminating conditions

First, the MLNN has four models referring to happy, sad, peaceful and angry categories. The timbre features for each song are computed to yield output decision values which determine the similarity attributes between the four models. During data training, 75 songs for every model category are fed into the MLNN classifier and the songs are classified into each model category. The net inputs into each unit in the hidden and output layers are shown in Equation (11).

$$I_j = \sum_i W_{ij} O_i + \theta_j \tag{11}$$

W_{ij} is the weight of unit i, from the previous layer j, O_i is the output of unit i, from the previous layer j and θ_j is a bias. The error is propagated backwards by updating the weights vector and biases to reflect the error of the network classification and the error is computed by Equation (12).

$$Err_j = O_j(1 - O_j) \sum_k Err_j w_{jk} \tag{12}$$

where w_{jk} is the weight of the connection from unit j to unit k in the next higher layer, and Err_k is the error of unit k.

The weight vector is updated in Equation (13), where w_i is the weight vector of the ith neuron, V is the number of input vector, μ is a learning rate initialized with big value at the beginning of the training and decreased with time and α is an activation function.

$$w_i(k+1) = w_i(k) + \alpha\mu(k)[V(k) - w_i(k)] \tag{13}$$

3. **Evaluation.** The experiment compares the performance of RENICA to other algorithms and from the results, it is verified RENICA improved the classification of music emotions. Using BSS Evaluation Toolbox [18] as shown in Equations (14), (15) and (16), the separation between the signal Source-to-Interference Ratio (SIR), Source-to-Distortion Ratio (SDR) and Source-to-Artifact Ratio (SAR) is computed. The quality of the separation is determined by higher values of SDR, SIR and SAR [6,7]. The observed separation qualities differ, when the powers of s_{target}, e_{interf}, e_{noise} and e_{artif} fluctuate across time.

3.1. **Source-to-Interference Ratio (SIR).** Interference may affect the quality of sound. In this paper, SIR is the source to interference ratio measured in dB which justifies the correlation between the original signal and estimated signal by measuring logarithm of original signal, s_{target} over the interfering sources e_{interf} as shown in Equation (14). The higher the SIR dB is, the better the separation is.

$$\mathrm{SIR} = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{14}$$

3.2. **Source-to-Distortion Ratio (SDR).** Signal-to-Distortion Ratio (SDR) was derived from the demixing components after SIR as measured in Equation (15), where s_{target} represents the original music signal, e_{interf} is the interference of the other sources, sensor noise is represented by e_{noise}, and e_{artif} indicates the artifact of the separation results.

$$\mathrm{SDR} = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{15}$$

3.3. **Sources-to-Artifacts Ratio (SAR).** As portrayed in Equation (16), SAR is used to measure the amount of artifacts presented by the separation algorithm, which in this paper represents the noise signal in musical components.

$$\mathrm{SAR} = 10 \log_{10} \frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2} \tag{16}$$

4. **Results and Discussion.**

4.1. **Music source separation performance.** We separate thousand music clips with 120-180 sec length (44.1 Hz) using RENICA. SIR, SDR and SAR are computed for both the singing voice and instrumental accompaniment as shown in Figure 3 and Figure 4 respectively. We compared the REPET, NMF and ICA separation models and later RENICA.

i) REPET: only using repeating pattern extraction techniques based on work in [6,7].
ii) NMF: only using nonnegative matrix factorization as done in [8].
iii) ICA-BSS: only using independent component analysis with blind source separation approach [10,11].
iv) RENICA-combination of REPET, NMF and ICA-BSS.

From Figures 3 and 4, the SIR, SDR, and SAR values for RENICA are higher in both singing voice separation as well as in instrumental accompaniment separation. In RENICA, the source separation minimized the error of the approximation of music mixture to sources estimation. For RENICA, SDR and SIR for both separations are significantly better than other models (SIR_{SV} = 26 dB, SIR_{IA} = 28 dB, SDR_{SV} = 18 dB, SDR_{IA} = 23 dB). Renica shows promising separation performance compared to the other state-of-the-art music source separation algorithms mainly because it represents all separation process, such as segment identification, sources estimation involving repeating and non-repeating estimation. The estimation then decomposes into singing voice and

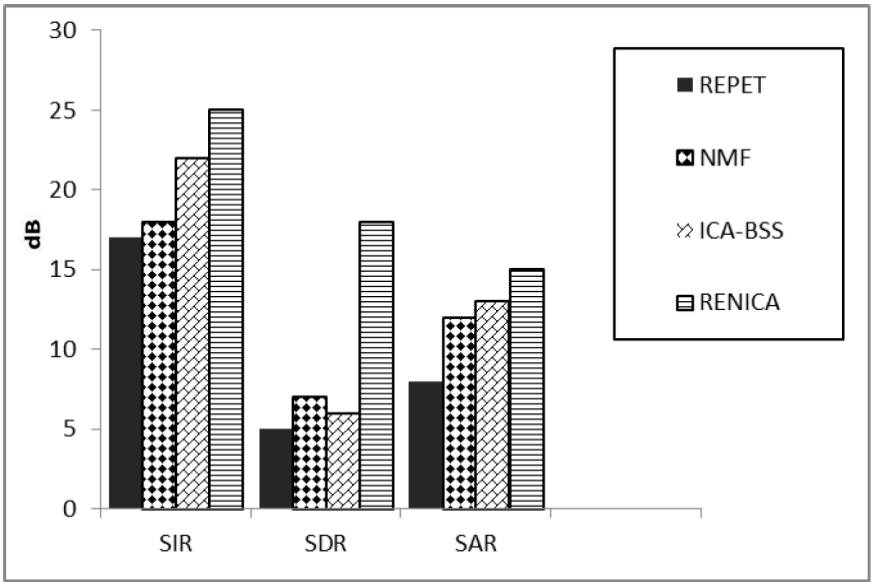

FIGURE 3. Singing voice separation

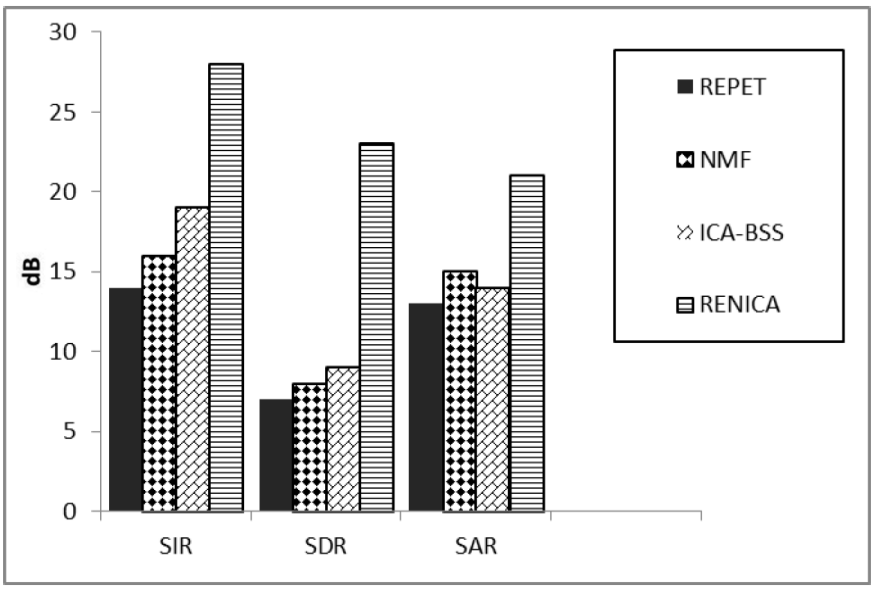

FIGURE 4. Instrumental accompaniment separation

instrumental accompaniment by finding and using coefficients or activation vector to find magnitude spectrogram of mixed signal as proposed in [8]. However, the interference ratio for ICA-BSS only varies about 7% from the proposed method compared to NMF 29% and REPET 45%. Overall the outcomes of music source separation are statistically significant for all techniques.

4.2. **Music emotion classification results.** We separated 1000 music signals into two parts: singing voice and instrumental accompaniment, resulting in 2000 music data. The datasets are then randomly mixed together and divided into 70% for training purposes and 30% for testing purposes. Based on the four model categories of music emotions, happy, sad, peaceful and angry – 75 songs for each model category are fed into the neural network classifier. The classification accuracies with and without RENICA are measured

for each category and the percentage of accuracy is derived from Equation (17).

$$\text{Accuracy} = \frac{\text{Correctly identified clips}}{\text{total number of testing clips}} \times 100\% \tag{17}$$

The testing results shown in Table 1 and Table 2 indicated the classification accuracy with and without RENICA respectively. From Table 1, by using RENICA the accuracy performance increases up to 97.33% while without RENICA, the accuracy performance is less effective with only 84.66%.

Table 1. Classification accuracy (with RENICA separation)

	Happy	Sad	Angry	Peaceful	Accuracy (%)
Happy	72	0	3	0	96.00
Sad	1	73	0	1	97.33
Angry	2	0	73	0	97.33
Peaceful	0	1	0	74	98.66
Total					97.33

Table 2. Classification accuracy (without RENICA separation)

	Happy	Sad	Angry	Peaceful	Accuracy (%)
Happy	58	4	13	0	77.33
Sad	2	63	0	10	84.00
Angry	10	0	65	0	86.66
Peaceful	0	7	0	68	90.66
Total					84.66

The sources of singing voice and instrumental accompaniment contain different timbre information leading to different emotions; thus, the separation process helps to select the best timbre information for music emotion classification.

5. **Conclusions.** This paper proposed RENICA for improving the separation of singing voice and instrumental accompaniment by combining three effective source separation algorithms namely REpeating Pattern Extraction Technique (REPET), Nonnegative Matrix Factorization (NMF) and Independent Component Analysis (ICA) with BSS approach. Higher values of SIR, SDR and SNR suggest better separation when using RENICA. Timbre features parameters extracted from singing voice and instrumental accompaniment have been successfully modelled and trained in ANN classifier. Our objectives and subjective evaluation suggest that separation of the music clips into SV and IA estimates using RENICA has improved automatic emotion classification and many improvements will be possible in the near future to improve its performance.

Acknowledgement. The authors would like to thank Universiti Malaysia Sarawak for supporting the research. Grant Number: F02(DPI28)/1244/2015(02).

REFERENCES

[1] A. Gabrielle and E. Stromboli, The influence of musical structure on emotional expression, *Music and Emotion: Theory and Research*, pp.223-243, 2001.

[2] A. S. Bhat, V. S. Amith, N. S. Prasad and D. M. Mohan, An efficient classification algorithm for music mood detection in western and Hindi music using audio feature extraction, *The 5th Int. Conf. Signal Image Process.*, pp.359-364, 2014.

[3] Y.-H. Chin, C.-H. Lin, E. Siahaan, I.-C. Wang and J.-C. Wang, Music emotion classification using double-layer support vector machines, *The 1st Int. Conf. Orange Technol.*, pp.193-196, 2013.
[4] R. Malheiro, R. Panda, P. Gomes and R. P. Paiva, Music emotion recognition from lyrics: A comparative study, *The 6th International Workshop on Machine Learning and Music (MML13)*, pp.9-12, 2013.
[5] J. Xu, X. Li, Y. Hao and G. Yang, Source separation improves music emotion recognition, *Proc. of International Conference on Multimedia Retrieval*, 2014.
[6] Z. Rafii and B. Pardo, Repeating pattern extraction technique (REPET): A simple method for music/voice separation, *IEEE Trans. Audio, Speech, and Language Processing*, vol.21, no.1, pp.73-84, 2013.
[7] K. Yoshii, M. Goto and H. G. Okuno, AdaMast: A drum sound recognizer based on adaptation and matching of spectrogram templates, *Proc. of the 5th Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, pp.184-191, 2004.
[8] A. Ozerov and C. Fevotte, Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3137-3140, 2009.
[9] C. Uhle, C. Dittmar and T. Sporer, Extraction of drum tracks from polyphonic music using independent subspace analysis, *Proc. of ICA*, pp.843-847, 2003.
[10] D. T. Pham, Mutual information approach to blind separation of stationary sources, *IEEE Trans. Information Theory*, vol.48, no.7, pp.1935-1946, 2002.
[11] J. Tang, W. Li and Y. Liu, Blind source separation of mixed PD signals produced by multiple insulation defects in GIS, *IEEE Trans. Power Delivery*, vol.25, no.1, pp.170-176, 2010.
[12] A. Klapuri, T. Virtanen and T. Heittola, Sound source separation in monaural music signals using excitation-filter model and EM algorithm, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.5510-5513, 2010.
[13] B. Rocha, R. Panda and R. P. Paiva, Music emotion recognition: The importance of melodic features, *The 5th International Workshop on Machine Learning and Music*, Prague, Czech Republic, no.2008, pp.1-4, 2013.
[14] F. Zheng, G. Zhang and Z. Song, Comparison of different implementations of MFCC, *Journal of Computer Science and Technology*, vol.16, no.6, pp.582-589, 2001.
[15] L. Muda, M. Begam and I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques, *arXiv:1003.4083*, 2010.
[16] J. Wang and K. Zhang, Music emotional classification and emotional curve fitting based on BP neural network, *International Conference on Computer Design and Applications (ICCDA)*, vol.2, 2010.
[17] C. Févotte, R. Gribonval and E. Vincent, *BSS_EVAL Toolbox User Guide – Revision 2.0*, 2005.
[18] S. Beveridge and D. Knox, A feature survey for emotion classification of western popular music, *Proc. of the 9th International Symposium on Computer Music Modeling and Retrieval*, pp.19-22, 2012.