Itakura-Saito Divergence Non Negative Matrix Factorization with Application to Monaural Speech Separation

A. Adewusi
Department of Electrical and Electronic Engineering, Federal University of Agriculture Abeokuta, Nigeria

K. A. Amusa
Department of Electrical and Electronic Engineering, Federal University of Agriculture Abeokuta, Nigeria

A. R. Zubair
Department of Electrical and Electronic Engineering, University of Ibadan

ABSTRACT
Monaural source separation is an interesting area that has received much attention in the signal processing community as it is a pre-processing step in many applications. However, many solutions have been developed to achieve clean separation based on Non-Negative Matrix Factorization (NMF). In this work, we proposed a variant of Itakura-Saito Divergence NMF based on source filter model that captures the temporal continuity of speech signal. The algorithm shows a very good separation results for mixture of two speech sources in terms of artifacts reduction. Besides that, Source to distortion ratio (SDR) and Source to Artifact Ratio (SAR) were found to be higher when compared with NMF algorithms with Kullback-Leibler and Euclidean divergences.

General Terms
Signal processing.

Keywords
Itakura-Saito divergence, monaural source separation, Non Negative Matrix Factorization,

1. INTRODUCTION
Noise and interference reduction is one of the core issues that signal processing field attempts to resolve. Particularly in audio processing, obtaining "clean" sounds is very critical to many applications such as speech recognition, speech decoding, music transcription, etc. Unfortunately in practice, noise/interference-free environment does not really exist. Generally, recorded sequences consist of a mixture of numerous sounds as input, while some are specified as target signals and the rest as noise depending on the area of application. In contrast, human listeners often have little trouble paying attention to a single source in the presence of conflicting multiple sources [11,5]. This is because the masking effect has made the human auditory system to respond asymmetrically to energy in the spectral domain. Then, it is logical to assume that metrically asymmetric cost functions, with similar error weighting characteristics to that of the human ear i.e. weighting over-estimation less than under-estimation, are more suitable for audio spectrogram factorization

Such mixtures are extremely commonplace; hence, this represents an important challenge.

Many solution approaches have been developed for monaural source separation challenge. One of such is to use the natural cues of the signal class to group the transform coefficients of the source signals in a sparse representation of their mixture. However, for speech mixture, this technique has been applied in different spectral domains. This is where features such as common onset/offset and amplitude/frequency co- modulation of a speech signal's time-frequency energy have been used to group localized segments of time-frequency energy. These are then merged using temporal structure such as pitch to solve separation challenge. Various Non-Negative Matrix factorization based techniques have been widely employed to achieve separation in this regime and these include shifted NMF [5]

Non-negative matrix factorization (NMF) has been proven to be well suited in the decomposition multivariate data [12], subject to non-negativity constraint especially in the blind audio source separation. Though optimal source separation algorithm is still a serious challenge, yet, a number of applications involving a human operator are starting to yield satisfactory results [6].

The problem is to estimate the sources that are present in a linear instantaneous mixture of M time-discrete input signals described by signal model:

$$S_m(n) = \sum_{m=1}^{M} S_n \text{ where, } 1 \le m \le M \ . \quad (1)$$

The mixture is thereafter transformed into a spectrogram X by the application of short-time Fourier Transform (STFT). Since the input signals $s_m(n)$ are real and the spectrogram is symmetric, while the part representing the negative frequency range can be ignored.

2. NON- NEGATIVE MATRIX FACTORIZATION
Nonnegative matrix factorization (NMF) of time-frequency representations such as the power spectrogram has become a popular tool in the signal processing community.

It has been exploited in single channel separation of sound signals, because it generates a parts-based decomposition of audio spectrograms where the parts typically correspond to individual notes or chords. Various single channel audio separation algorithms based NMF have been developed. For example, perpetually weighted NMF was applied on the power spectrogram in order to model the human auditory system for sound attributes characterization as in [11]. In addition to that, a NMF model in Bayesian domain for monaural source separation has been proposed in [13], However, a prominent disadvantage of NMF is the task of performing the clustering of the basis functions to their individual sources after decomposition. With continual

progress that have been made in the search for algorithms for clustering the basis functions to sources, it still remain an open area of research in which much works are required to developed optimal clustering algorithms.

Given such a time-frequency representation $V \in R_{+}^{F\times N}$, NMF consists in finding a factorization of the form $X \approx BG$

where $B \in \mathbb{R}_{+}^{F\times N}$, $G \in \mathbb{R}_{+}^{K\times N}$ and $K \ll F, N$.

The factorization is obtained by minimizing a cost function of the form $D(\mathrm{X}, \mathrm{BG})$. NMF as expressed in [8] approximates a nonnegative, real-valued matrix X of size $K\times T$ as a product of matrices B and G .

$$X \approx \tilde{X} = BG \tag{2}$$

where B is of size $F\times N$ and G is of size $K\times N$, with I as user defined parameter. The process of estimating B and G is an optimization problem, the sole aim of which is to minimize some cost function, $C(\cdot)$, with respect to G and B, subjected to non-negativity constraint on G and B, which is expressible in the form written as:

$\{B, G\} = \arg\min \mathrm{C}(A; B, G)$ [3], provided $B, G \geq 0$.

In the areas of speech and audio applications, where NMF is normally applied to spectrogram data, the cost functions are selected based on their perceptual relevance. In this work, Itakura Saito (IS) distance NMF algorithm is applied for the decomposition of the monaural mixture.

From the foregoing, applying the IS cost function to minimize the problem:

$$C_{IS} = (A \setminus BG) = \left| \frac{A}{BG} - \ln\left[\frac{A}{BG}\right] - 1 \right| \tag{3}$$

The preceeding cost functions are all generalized by β divergence in [3] , such that:

$$C_{\beta}(A \| BG) = \sum A_{i,j} \frac{A_{i,j}^{\beta-1} - (BG)_{i,j}^{\beta-1}}{\beta(\beta-1)} - \frac{(BG)_{i,j}^{\beta-1}}{\beta} \frac{(BG)_{i,j} - A_{i,j}}{\beta} \tag{4}$$

Minimizing the cost function in [11]

$$C(B, G) = C_r(B, G) + \alpha C_t(G) + \beta C_s(G) \tag{5}$$

where the reconstruction error $c_r(B, G)$ and also the temporal continuity

$$c_t(G) = \sum_{j=1}^{J} \frac{1}{\sigma^2} \sum_{t=2}^{T} \left(g_{t,j} - g_{t-1,j} \right)^2 \tag{6}$$

while the minimization is obtained as:

$$\sigma_j = \sqrt{\frac{1}{T} \sum_{t=1}^{T} g_{t,j}^2} \tag{7}$$

The sparseness criterium $C_s(G)$ is given as:

$$C_s(G) = \sum_{j=1,t=1}^{J,T} f\left(g_{j=1,t=1} / \sigma_t \right) \tag{8}$$

Initialization of the matrices B and G are performed with the absolute values of Gamma noise (IS), the multiplicative update rules are thereafter applied in order to solve the optimization problem.

2.1 Itakura- Saito divergence and its scale invariance property

This divergence was derived by Itakura and Saito (1968) from the Maximum Likelihood (ML) estimation of short-time speech spectra under autoregressive modeling. It was defined as measure of the goodness of fit between two spectra and became popular in the speech community during the seventies. The divergence was applauded for its good perpetual quality of reconstructed signals.

Authors in [7] developed a variant of an unsupervised inference procedure for audio source separation. Here, components in nonnegative matrix factorization (NMF) are clustered automatically in audio sources with the aid of a penalized maximum likelihood scheme. The penalty term that was imposed on this separation favored group sparsity, this is prompted by the assumption that the local amplitude of the sources are statistically independent. The algorithm extends multiplicative updates for NMF which leads to the proposition of a test statistic to tune hyper-parameters.

For music transcription [1] proposed an Itakura Saito NMF based tempering approach convergence of IS-NMF to global minima. This algorithm is based on NMF with the beta-divergence, where the shape parameter beta acts as a temperature parameter.

In [6] is the adaptation of spectrum dictionaries in audio source separation with supervised learning. Here it was assumed that sample data of the audio sources to separate are available, a filter adaptation in the frequency domain is presented in the Non-Negative Matrix Factorization with the Itakura-Saito divergence regime. The algorithm had a capacity to retrieve the acoustical filter applied to the sources with high level of accuracy and demonstrated tangible higher performances on separation tasks with respect to the non-adaptive model.

One of the unique properties of IS divergence is that it is scale invariant, meaning that low energy components of X bear the same relative importance as high energy ones. This is relevant to situations in which the coefficients of X have a large dynamic range, such as in audio short-term spectra. The IS divergence equally results to desirable statistical interpretations of the NMF problem. Indeed, NMF can be recast as maximum likelihood (ML) estimation of B and G in superimposed signals under simple Gaussian assumptions [4]. Additionally, IS-NMF can be interpreted as Maximum

Likelihood of B and G in multiplicative Gamma noise [3], [4].

The IS divergence belongs to the class of Bregman divergences and has a limit case of the β divergence. Thus, the gradient descent multiplicative rules in these condensed form:

$$d\beta = (\gamma_x / \gamma_y) = \gamma^B d_\beta (x/y) \quad (9)$$

Which by implication makes IS divergence scale-invariant, that is $d\beta = (\gamma_x / \gamma_y) = d_{IS}(x/y)$, the only one with this unique property among the β divergence family . The scale invariance means that the same relative weight is assigned to small and large coefficients of X in cost function such that a bad fit of the factorization for a low-power coefficient $[X]_{fn}$ will cost as much as a bad fit for higher power coefficient $[X]_{f^i n^i}$. On the other hand, factorizations obtained with $\beta > 0$ (such as with the Euclidean distance or the KL divergence) will rely more heavily on the largest coefficient so that less accuracy is expected in the recovery of the low-energy components.

2.2 Computation of the update rules

Estimation of the basis function B and the gain G of the j^{th} basis function in frame T is carried out by applying the following update rules on vectors B and G respectively.

$$B \leftarrow B \cdot \frac{XG^T}{BGG^T} \quad (10a)$$

$$G \leftarrow G \cdot \frac{B^T X}{B^T BG} \quad (10b)$$

Minimizing (IS) using the gradient criterion $D_B(X / \mathrm{BG})$ with respect to B and G so that from [4]:

$$\nabla_G D_\beta(\mathrm{X/\ BG}) = \mathrm{B}^T \left((BG)^{\beta-2} \right) \cdot (BG - X) \quad (11)$$

$$\nabla_B D_\beta(\mathrm{X/\ BG}) = \left((BG)^{\beta-2} (BG - X) \right) G^T \quad (12)$$

where $(\cdot)$ denotes the Hadamard entry wise products and $A^{\cdot \mathrm{n}}$ denotes the matrix with entries $[A]^n_{i,j}$. The multiplicative gradient descent technique is equal to updating each parameter by multiplying its successive value obtained in previous iteration by the ratio of the negative and positive parts of the derivative of the criterion with respect to this parameter which are:

$$\theta \leftarrow \theta \cdot [\nabla f(\theta)]_- / [\nabla f(\theta)]_+ \quad (13)$$

$$\nabla f(\theta) = [\nabla f(\theta)]_- / [\nabla f(\theta)]_+ \quad (14)$$

Applying these, guarantee the non-negativity of the parameter updates, if and only if the initialization is setup with a nonnegative value. So that the update rules are now formed as:

$$B \leftarrow B \cdot \frac{\left((BG)^{\beta-2} \cdot X \right) G^T}{(BG)^{\beta-1} G^T} \quad (15)$$

$$G \leftarrow G . \frac{B^T \left((BG)^{.\beta-2} . X \right)}{(BG)^{.\beta-1} B^T} \quad (16)$$

Alternatively, it can be expressed as a variant of Bregman divergence which takes the form of:

$$d\phi\left(x/y\right) = \phi(x) - \phi(y)(x - y) \quad (17)$$

where ϕ is a strictly convex function of ∇ that has a continuous derivative $\nabla\phi$. The IS divergence is obtained with $\phi(y) = -\log(y)$ as in [4]. This gives:

$$B \leftarrow B \cdot \frac{\left(\nabla^2 \phi(BG)^{.\beta-2} \cdot X \right) G^T}{\left(\nabla^2 \phi(BG) \cdot BG \right) G^T} \quad (18)$$

$$G \leftarrow G \cdot \frac{B^T \left(\nabla^2 \phi(BG) \cdot X \right)}{B^T \left(\nabla^2 \phi(BG) \cdot BG \right)} \quad (19)$$

3. ITAKURA-SAITO MULTIPLICATIVE UPDATE ALGORITHM

In the minimization of the NMF problem, IS divergence is imposed on the factorization stage of the source filter model in [8]. This is done in order to further improve the separation performance and to investigate the computational efficiency (convergence and run time) of IS algorithm over other types of cost functions earlier applied on the model.

3.1 IS-multiplicative update gradient descent

This is perhaps the most efficient technique of evaluating the matrices B and G through gradient descent by multiplicative updates. This update approach exploits the fact that multiplying any two nonnegative values produces another non-negative value. Then, initializing the elements of G and B to non-negative values and given a non-negative **A**, the non-negativity constraint is imposed by applying multiplicative updates to B and G ; this also implies that if an element of either factor is assigned the value zero, during the update it remains at zero. One interesting characteristics of this technique is that it can be extended to a large number of cost functions. The IS algorithm is thus derived by setting $\beta = 0$ and $\phi y = -\log(y)$.

This gradient descent algorithm has a normalization step at each iteration and this prevents trivial scale indeterminacies, thereby returning a constant valued cost function [4].

Summary of the IS-NMF Algorithm

Input: Non-Negative Matrix X

Output: Non Negative Matrices B and G such that

$$X \approx BG$$

Initialize: B and G with non-negative values

for $i = 1:n$ do

$$B \leftarrow B \cdot \frac{\left((BG)^{\cdot(-2)} \cdot X\right) G^T}{(BG)^{\cdot(-1)} G^T}$$

$$G \leftarrow G \cdot \frac{B^T \left((BG)^{\cdot(-2)} \cdot X\right)}{(BG)^{\cdot(-1)} B^T}$$

Normalize B and G

End

4. SEPARATION

In the estimation of speech mixture, a model based monaural source separation algorithm is applied. Here a linear instantaneous mixture will be considered with a single gain from equation (1), which lead to model that corresponds to a source filter model.

Non negative Matrix Factorization separates the magnitude spectrogram X of the mixture into a number of I channels, each with respective spectrograms $c_i = 1 \le i \le I$. The main motivation for adopting the NMF for the source separation problem is that it captures the pitch structure of speech and much more in its spectrogram representation. Inasmuch as human speech utterance and acoustic musical notes are pitched in nature, they can be described by a constant frequency basis vector B_i and a time varying gain G_i which equals envelope of a single note. NMF Mel Frequency Cesptral Coefficient (NMF-MFCC) clustering model as in [8] was applied to cluster the basis functions that belong to the same source.

The i-th column of B and the i-th row of G when multiplied form the spectrogram c_i of the i-th channel

$C_i = B_i G_i$ where C_i is of rank one matrix [8] .

It is assumed that the row G_i of matrix G is low pass due to the continuity attribute of acoustic signals. It has been established that an additional cost function C_t is required when considering temporal continuity which tends to improve the separation quality for NMF algorithms particularly in speech separation.

Thus, this cost function is expressed as:

$$c_t = a \sum_i \frac{\sum_{t=2}^{T} \left(G(i,t) - G(i,t-1)\right)^2}{\sum_{t=1}^{T} G^2(i,t)} \qquad (21)$$

where T is the drop factor, since the mixture consists of sources of different length.

5. EXPERIMENTAL SETUP

The algorithm is implemented in MATLAB for single channel audio mixtures. A male voice and a female voice are recorded at Phonology Laboratory, University of Ibadan, Nigeria. Each of the speakers are made to make a sentence of which varied in duration of roughly 4 to 8 seconds at a sampling frequency of 16 *kHz*

The magnitude spectrogram of the time-domain signal was obtained using the STFT. Hann windows of 4096 samples in length were selected, while 75% overlap was allowed between the successive Hann windows. The number of NMF basis functions (channels) for the test signals were equal to 15. The number of frequency basis functions may vary with the length (time duration) of the test samples in the test set used. In this work NMF was run for up to 600 iterations.

Matrices B and G are randomly initialized with non-negative values. The cost function employed in the decomposition of the mixture is IS . The multiplicative updates and positive initialization for B and G ensure the factorization is non-negative. The algorithm is set for number of sources equal to 2 and it runs for 600 iterations. The number of translations k is 15.

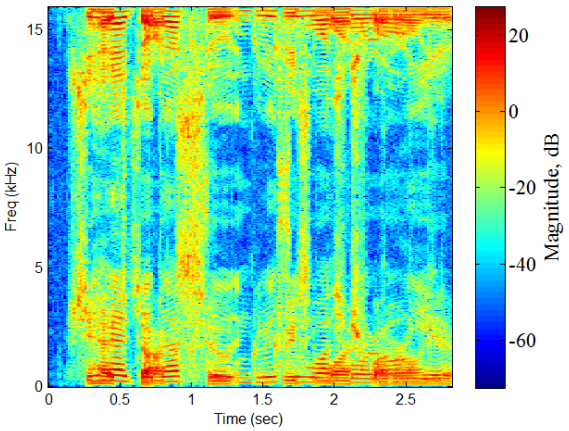
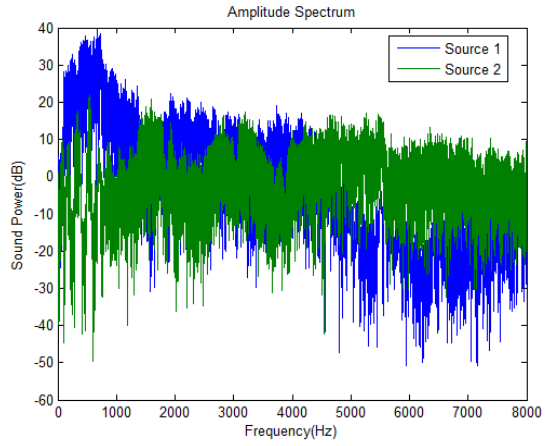


Figure 2: Spectrogram of the Mixture are NMF

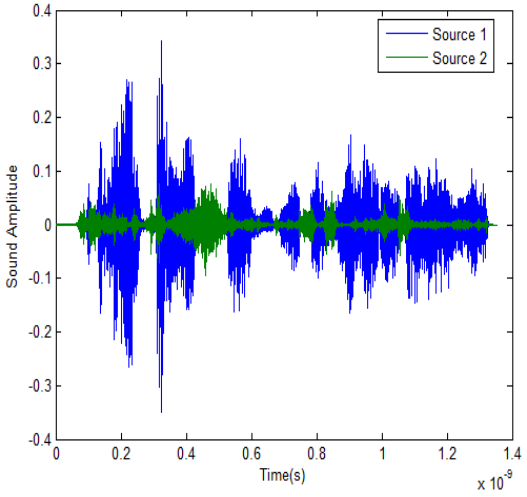
The figure above shows the decomposition of the speech mixture after decomposition by IS-NMF algorithm. The high energy region (0-20dB) shows the distribution of the formants that represent the components of the mixture.

5.1 Separation Results

It has been earlier reported in the methodology that IS-NMF was used in the data decomposition stage of this separation based on its excellent performance.



(*a*)



(*b*)

Figure 3: Plots of separated sources in different domains (a) frequency (b) time

Figure 3 (a) shows the amplitude spectrum of the separated sources. At the low end of the spectrum, source 1 has the highest spectral power and this reduces as it moves along the frequency spectrum. This corresponds to the male voice which is characterised with high pitch at low frequency. Source 2 shows an increasing trend in amplitude as it moves along the frequency spectrum. This corresponds to a female voice.

From Figure 3(b) the time domain plot of the separated mixture is shown; the gap in between the sources shows the effect of sparsity which is a very important attribute of speech signal.

5.2 Performance Evaluation

Performance metrics were computed to assess the robustness of the separation model. The result obtained using $IS\text{-}NMF_{mfcc}$ is compared with other factorisation models (Euclidean and Kullback - Leibler divergences) within the dynamic range of $5dB$ to $40dB$. The separation results for M = 2 is presented in Table 1.

Table 1: Performance Metrics for the Separation Algorithm

R	*IS - NMF*		*KL – NMF*		*EU – NMF*	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**
SDR						
5	10.97	12.66	10.18	8.73	8.71	10.02
10	12.21	11.09	8.44	9.79	6.36	7.96
20	9.39	11.15	8.24	9.21	8.89	7.58
40	-10.31	-13.70	-13.70	-9.10	-13.84	-8.62
SAR						
5	10.97	11.66	9.12	-8.73	-7.71	10.02
10	8.22	11.09	8.44	9.80	4.36	5.96
20	9.39	10.15	6.24	9.21	8.89	4.58
40	-12.17	-10.31	-13.70	-9.10	-13.84	-8.62

Note: S1 and S2, respectively, represent source 1 and source 2; R denotes the dynamic range in dB

From Table 1, it is found that the Source to Distortion Ratio (SDR) and Source to Artefact Ratio (SAR) of $IS\text{-}NMF$ are the highest among the three separation models. Both computed values of SAR and SDR increase steadily over the considered range. With this result, the performance of $IS\text{-}NMF$ is better than others over the dynamic range of mixtures. The most probable reason is that the divergence is not easily distorted by small values of noise when compare with other two models.

Itakura-Saito distance has been found to be the most suitable cost function for the decomposition of non-negative data mixture, as it converges to local minimum over any number of iterations. Besides this, increasing the number of channels or the size of K yields better results of iteration and ultimately improved the quality of factorization.

Moreover, the model produced a good separation results with high values of SDR and SAR for each of the separated sources. This is evident in Table 1, where the computed data show that the performance of the $IS\text{-}NMF$ algorithm is highest among the NMF algorithms considered.

6. CONCLUSION AND FUTURE WORKS

We have constructed an Itakura –Saito divergence NMF for source separation which is applied in the estimation of monaural speech mixture. The algorithm achieved a very high separation result compared to other divergence-based algorithms in terms of SDR and SAR. From the foregoing, the algorithm has the potential of separating speech mixture that contains more than two sources. However, eliminating the constraint that the local minimal introduces into the cost function of the NMF is still an open issue, a probable solutions will be to adopt variants of Markov chain Monte Carlo (MCMC) simulation scheme to provide a deeper insight on the order of the criteria to achieve the minimization.

Also, the possibility of implementing Kalman filter on the separated outputs in order to minimize the artifacts that may be present in the estimated sources can be explored.

This model can be extended, to other audio applications such as speech recognition, pitch modification and automatic music transcription. These applications would benefit from the availability of segregated sound sources from the mixture of audio signals to aid further processing.

7. REFERENCES

[1] Bertin N, F ´ Evotte C, Badeau R. 2012: A Tempering Approach For Itakura-Saito Non-Negative Matrix Factorization. With Application To Music Transcription. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Signal Processing Society 2012 Kyoto Japan

[2] Bensaid, S. S. 2009 Monomicrophone Blind Audio Source Separation Using Thee-M Kalman Filter And Short+ Long Term AR Modelling. IEEE *Conference On Signals and Systems* (Pp. 343 - 345;). Califonia:.

[3] Cahill, N. M. 2012 : An Investigation Of The Utility Of Monaural Sound Source Separation Via Nonnegative Matrix Factorization Applied To Acoustic Echo And Reverberation Mitigation For Hands-Free Telephony. Doctoral Thesis. Callan Institute, Department Of Electronic Engineering.

[4] Fevotte, C. Bertin, N.and Dirrieu, J.L. 2009 Nonnegative Matrix Factorization With The Itakura-Saito Divergence. With Application To Music Analysis. Neural computations Vol 21, No 3 793-830

[5] Jaiswal, R. 2013 Non-Negative Matrix Factorization Based algorithms To Cluster Frequency Basis Functionsalgorithms. Doctoral Thesis. Dublin Institute Of Technology, School Of Electrical Engineering Systems: Dublin Institute Of Technology

[6] Jaureguiberry X, Leveau. P, Maller. S, Jos´E Burred . J. 2011. Adaptation Of Source-Specific Dictionaries In Non-Negative Matrix Factorization For Source Separation. IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP). IEEE Signal Processing Society. Prague, Czech Republic 1-4

[7] Lef`Evre, A. Bach, Y.F and F´Evotte, C. 2011. Itakura-Saito Nonnegative Matrix Factorization with Group Sparsity Prague, Czech Republic. ICASSP May 22-27 IEEE International Conference on Acoustics, Speech nd Signal Processing (ICASSP). IEEE Signal Processing Society. Prague, Czech Republic.

[8] Martin Spiertz, V. G. 2009. Source-Fiilter Based Clustering For Monaural Blind Source Separation. *12th Int. Conference On Digital Audio Effects* Italy 1-7.

[9] Mikkel, S. 2008. Single-channel source separation using non-negative matrix. Technical University of Denmark, Informatics and Mathematical Modeling. Denmark: Technical University of Denmark

[10] Nobutaka Ono, K. M. 2008. Separation Of A Monaural Audio Signal Into Harmonic/Percussive Components By Components By Complementary Diffusion on Spectrogram. European *Signal Processing Conference* . Paris: European Signal Processing Society 1-4.

[11] Virtanen. T. O 2006. Monaural Sound Source Separation By Perceptually *Weighted Non-Negative Matrix Factorization. IEEE* Transactions *On Signal Processing*, 1-8

[12] Lee, D. D. and Seung, H. S. 2001 : Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing,* , MIT Press vol. 13, 1-7

[13] Chien J.T and Yang P.K 2016**;** Bayesian Factorization and Learning for Monaural Source Separation: IEEE/ACM Transactions on Audio, Speech, And Language Processing, Vol. 24, No. 1, 185-195