A PERCEPTUALLY ENHANCED BLIND SINGLE-CHANNEL AUDIO SOURCE SEPARATION BY NON-NEGATIVE MATRIX FACTORIZATION *

S. Kırbız and B. Günsel

Multimedia Signal Processing and Pattern Recognition Lab.,
İstanbul Technical University, Dept. of Electronics and Communications Engineering
34469 Maslak İstanbul, Turkey
{kirbiz,gunselb}@itu.edu.tr

ABSTRACT

This paper proposes a 2D Non-negative Matrix Factorization (NMF) based single-channel source separation algorithm that emphasizes perceptually important components of audio. Unlike the existing methods, the proposed scheme performs a psychoacoustic pre-processing on the mixture spectrogram in order to supress audio components that are not critical to human hearing sensation while amplifying the perceptually important ones. This yields the auditory spectrogram referred as sonogram of the observed audio mixture and the individual sources are then extracted by 2D NMF. Test results reported in terms of Signal-to-Distortion-Ratio (SDR), Signal-to-Inference-Ratio (SIR) and Signal-to-Artifact-Ratio (SAR) show that the proposed perceptually enhanced separation improves the quality of decomposed audio sources by 1.5-6.5 dB with a reduced computational complexity.

1. INTRODUCTION

Estimating individual audio sources from the mixture signal is called audio source separation. Audio source separation has been used in several applications including robust speech recognition, music transcription, speaker identification, etc.

In this paper the focus is on the separation of music and speech signals from a single observation. The goal of single-channel Blind Source Separation (BSS) is to extract the underlying audio source signals from a single linear mixture. In such situation, human listener has the ability to keep the attention to a single audio source in an adverse acoustical condition. However, the problem of estimating several sources from one input signal is an ill-posed problem thus has been a challenging topic for the researchers.

A vast amount of research has been conducted in the field of blind single-channel audio source separation. Among these, Non-negative Matrix Factorization (NMF) [1, 2, 3] is a simple but efficient factorization method which has been extensively used for factorizing the input data into a linear combination of basis vectors with non-negativity constraints on output matrices. Efforts have been made to develop more robust and efficient algorithms by adding further constraints for the decomposition, such as sparseness, temporal continuity [4] or extending the model to be convolutive [3, 2]. Smaragdis [3] introduced the non-negative matrix factor deconvolution (NMFD) algorithm in which each instrument is modeled by a time-frequency signature that varies in intensity over time. Scmidt et al. [2] proposed non-negative matrix factor 2-D deconvolution (NMF2D) algorithm for separating the instruments in polyphonic music by representing each instrument by a single time-frequency profile convolved in both time and frequency in a log-frequency spectrogram. A few approaches in the area of source separation have utilized the framework of psychoacoustics [5]. Among these, Virtanen [1] presents a perceptually weighted NMF algorithm for single channel source separation that assigns a weight coefficient for each critical band in each frame to model the loudness perception of the human auditory system. Although the algorithm achieves a high separation quality on non-overlapping audio sources, it is not sufficient for separating the mixtures of instruments/sources which overlap their whole duration. In [6], a perceptually motivated Frequency-Domain Independent Component Analysis (FDICA) scheme is proposed which filters the frequency components that are perceptually irrelevant by exploiting the masking properties of speech. The deficiency of the frequency domain algorithms is the invariance to scaling and permutation which means that the output of the separation algorithm will be the original sources, arbitrarily scaled, permuted and delayed. In this work, the psychoacoustic masking applied prior to FDICA is used to avoid the permutation problem rather than improving the perceptual quality of the separated sources.

In the proposed method, a Non-negative Matrix Factor 2-D Deconvolution (NMF2D) [2] based perceptual source separation is performed by applying a psychoacoustic pre-processing prior to decomposition. The pre-processing is applied based on the psychoacoustic model proposed in [7] in order to remove the information in the audio signal which is not critical to our hearing sensation while retaining the important parts. In [7], the raw audio signals are pre-processed in order to obtain a time-invariant representation of the perceived characteristics in two stages. In the first stage of the feature extraction process, the specific loudness sensation (sone) per critical band (Bark) is calculated. In the second stage, the periodicity and the spectrum histograms are calculated based on the pre-processing and combined with the meta-information. The extracted features are then clustered and organized on a 2D map display using Self Organizing Maps. We just applied the psychoacoustic pre-processing

* This work is partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) BIDEB.

step of the work proposed in [7] in order to increase the perceptual quality of the separated sources. We evaluated the performance of the proposed method on real audio mixtures which were synthetically generated by summing two different sources of music and speech. The effects of the NMF2D parameters are also investigated by performing the tests using combinations of various NMF2D parameters on our dataset which are defined in Section 3. It is shown that the proposed method enhances the separation performance by an amount of 1.5 dB to 6.5 dB and outperforms the conventional NMF [8] and NMF2D [2] algorithms.

2. PROPOSED METHOD

Conventionally Non-negative Matrix Factorization (NMF) based source separation algorithms aim to factorize the observed non-negative mixture matrix $\mathbf{X} \in \mathbb{R}^{B\times M}$ as a product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$X[b,i] \approx \sum_{n=0}^{N-1} W[b,n]H[n,i] \tag{1}$$

where $W[b,n]$ denotes the $b-$th basis component of the $n-$th audio component, $H[n,i]$ is the gain of the $n-th$ component in time frame $i = 0\cdots M-1$ and N is the number of audio components. The audio components belonging to the same source are then clustered into a single source.

In audio source separation applications, the mixture $\mathbf{X}$ is usually represented as a time-frequency spectrogram. Given the observed mixture $\mathbf{X}$, we are interested in jointly estimating the basis ($\mathbf{W} \in \mathbb{R}^{B\times N}$) and the gain ($\mathbf{H} \in \mathbb{R}^{N\times M}$) matrices which are restricted to be entry-wise non-negative. The rank N of the factorization is usually chosen such that $(B+M)N < BM$, and hence the dimensionality reduction is achieved.

Unlike the existing methods, the proposed perceptually enhanced NMF2D framework yields a clustered representation of the mixture data by performing a psychoacoustic pre-processing on the spectrogram. Hence the proposed method improves the quality of the separated sources as well as decreases the computational complexity. In this section, we first describe the psychoacoustic pre-processing applied on the mixture spectrogram and then present the proposed decomposition method.

2.1. Auditory Spectrogram

Most of the audio source separation algorithms perform the separation on audio spectrogram. In this work, we propose to apply a pre-processing scheme on the spectrogram in order to retain the perceptually important components and supress the information which is not critical to human hearing. The pre-processing is applied using the MA Toolbox [7]. First, the Short-Time Fourier Transform (STFT) of the mixture signal $\mathbf{x}$ is computed

$$X_F[k,i] = \frac{1}{N_F}\sum_{t=0}^{N_F-1} h[t,N_F]x[t,i]e^{-j2\pi tk/N_F}, \tag{2}$$

where $0 \leq k \leq \frac{N_F}{2}$, $k = 0\cdots K-1$ is the frequency index, $i = 0\cdots M-1$ is the frame index, t is the sample index, N_F is the frame length and h is the Hanning window. To model the frequency response of the outer and middle ear, each frequency component of the spectrogram is weighted as

$$|S_w[k,i]|^2 = W_V^2[k]|X_F[k,i]|^2, \tag{3}$$

where the weighting function is defined as

$$W_V[k] = 10^{A_{dB}\left(\frac{kF_s}{N_F}\right)/10}. \tag{4}$$

In Eq.(4), F_s is the sampling frequency and $A_{dB}(.)$ is the response of the outer and middle ear model to each frequency f (kHz), proposed by Terhardt [9]:

$$\begin{aligned} A_{\mathrm{dB}}(f_{\mathrm{kHz}}) = &-3.64(10^{-3}f)^{-0.8} - 10^{-3}(10^{-3}f)^4 \\ &+ 6.5\exp\big(-0.6(10^{-3}f-3.3)^2\big). \end{aligned} \tag{5}$$

The main characteristic of this weighting filter is that the influence of very high and low frequencies is reduced while frequencies around $3-4$kHz are emphasized [7].

Subsequently the frequency bins of the STFT are grouped into 24 critical-bands according to [5] in order to obtain the spectrogram in bark scale represented as $C[b,i]$, where $b = 0\cdots 23$ is the critical band index and $i = 0\cdots M-1$ is the frame index. The conversion between the bark and the linear frequency scale is computed with,

$$Z_{\mathrm{bark}}(f_{\mathrm{kHz}}) = 10\arctan(0.76f) + 3.5\arctan(f/7.5)^2. \tag{6}$$

These frequency bands reflect the characteristics of the human auditory system. The width of the critical-bands is linear from 100Hz to 500Hz and beyond 500Hz the width increases nearly exponentially [5].

We apply a spectral masking on the bark spectrogram according to [7]

$$S_m[b,i] = \sum_{p=0}^{23} 10^{T_{\mathrm{dB}}[b,p]/10} C[p,i] \tag{7}$$

where $T[b,p]$ is the contribution of critical-band z_b to z_p

$$\begin{aligned} T[b,p] =&15.81 + 7.5(z_p - z_b + 0.474) \\ &- 17.5\big(1 + (z_p - z_b + 0.474)^2\big)^{1/2}. \end{aligned} \tag{8}$$

The main characteristic is that lower frequencies have a stronger masking influence on higher frequencies than vice versa.

The specific loudness sensation (sone) is calculated using the formula suggested by Bladon and Lindblom [7] in decibel relative to the threshold of hearing,

$$X[b,i] = \begin{cases} 2^{(S_{m_{dB}}[b,i]-40)/10}, & \text{if} S_{m_{dB}}[b,i] \geq 40\text{dB}, \\ (S_{m_{dB}}[b,i]/40)^{2.642}, & \text{otherwise}, \end{cases} \tag{9}$$

where $S_{m_{dB}}$ is the loudness in dB defined as

$$S_{m_{dB}}[b,i] = 10\log_{10} S_m[b,i]. \tag{10}$$

Fig. 1 illustrates an audio signal representation in time and various representations in time-frequency domain.

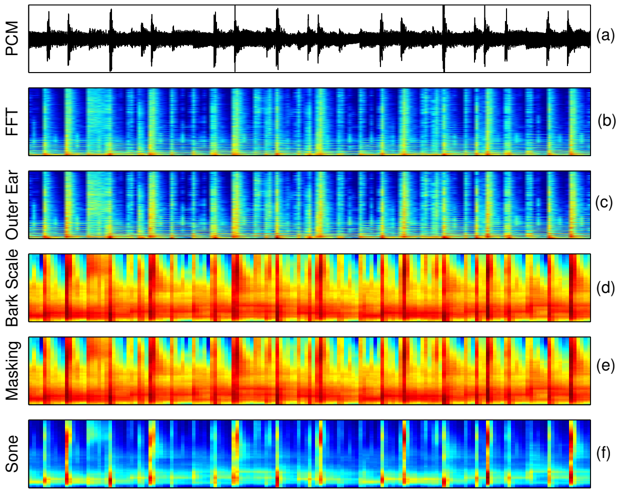

Fig. 1. The time and time-frequency illustrations of an audio signal. (a) time domain signal $\mathbf{x}$ (b) spectrogram $\mathbf{S}$, (c) filtered spectrogram $\mathbf{S}_w$ (d) Bark spectrogram $\mathbf{C}$ (e) masked Bark spectrogram $\mathbf{S}_m$, (f) sonogram $\mathbf{X}$.

Fig.1(a) depicts the signal in time domain. In Fig.1(b), the audio spectrogram $\mathbf{S}$ is displayed. Fig. 1(c) is a plot of $\mathbf{S}_w$ obtained after applying outer and middle ear filtering to the spectrogram $\mathbf{S}$. Fig.1(d) illustrates the Bark spectrogram $\mathbf{C}$ obtained by grouping the frequency bands in $\mathbf{S}_w$ into critical bands. Fig. 1(e) is $\mathbf{S}_m$ obtained after frequency masking applied on the Bark spectrogram $\mathbf{C}$ and Fig. 1 (f) illustrates the auditory spectrogram (sonogram) $\mathbf{X}$ obtained after loudness calculation.

As it is seen from Fig.1, the sonogram retains the auditory components which are critical to human hearing. Note that, since the sonogram has only $B = 24$ frequency bands, sonogram representation significantly decreases the size of the time-frequency data which consequently decreases the computational complexity of decomposition.

2.2. Source Decomposition by Non-negative Matrix Factor 2-D Deconvolution

In Non-negative Matrix Factor 2-D Deconvolution (NMF2D) based source separation algorithms [2], the observed mixing data $\mathbf{X} \in \mathbb{R}^{K \times M}$ is factorized to be a 2-D convolution of $\mathbf{W}^{\tau}$ which depends on time τ and $\mathbf{H}^{\phi}$ which depends on pitch ϕ:

$$\mathbf{X} \approx \mathbf{\Lambda} = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{\Phi-1} \overset{\downarrow\phi}{\mathbf{W}^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}. \tag{11}$$

In (11), $\downarrow \phi$ denotes the downward shift operator which moves each element in the matrix ϕ rows down, and $\rightarrow \tau$ denotes the right shift operator which moves each element in the matrix τ columns to the right [2]. Each element in Λ is defined as:

$$\mathbf{\Lambda}[b,i] = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{\Phi-1} \sum_{n=0}^{N-1} W^{\tau}[b-\phi,n] H^{\phi}[n,i-\tau], \tag{12}$$

where N is the number of audio components. Note that, NMF model is a special case of NMF2D model for $\tau = 0, \phi = 0$.

In the literature, NMF algorithms [8] are used to estimate the non-negative basis functions and mixing matrices iteratively based on the minimization of the Euclidean distance between the observed data $\mathbf{X}$ and model $\mathbf{\Lambda}$, or divergence D, given as

$$D = \sum_{b} \sum_{i} X[b,i] \log \frac{X[b,i]}{\Lambda[b,i]} - X[b,i] + \Lambda[b,i], \tag{13}$$

where $b = 0 \ldots B-1$ is the frequency index and $i = 0 \ldots M-1$ is the frame index. In most of the NMF based algorithms, the frequency index is the index of the linear frequency bands. In the proposed method, the linear frequency bands $k = 0 \cdots K - 1$ are grouped into critical bands $b = 0 \cdots B - 1$ as described in [7].

Considering the gradient decent optimization scheme, the multiplicative updates for $\mathbf{H}^{\phi}$ and $\mathbf{W}^{\tau}$ are obtained as:

$$\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} .* \left[\sum_{\tau} \overset{\downarrow\phi}{\mathbf{W}^{\tau}}{}^{T} \left(\overset{\leftarrow\tau}{\mathbf{X}} ./ \overset{\leftarrow\tau}{\mathbf{\Lambda}} \right) \right] ./ \left[\sum_{\tau} \overset{\downarrow\phi}{\mathbf{W}^{\tau}}{}^{T} \mathbf{1} \right] \tag{14}$$

$$\mathbf{W}^{\tau} \leftarrow \mathbf{W}^{\tau} .* \left[\sum_{\phi} \left(\overset{\uparrow\phi}{\mathbf{X}} ./ \overset{\uparrow\phi}{\mathbf{\Lambda}} \right) \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}{}^{T} \right] ./ \left[\sum_{\phi} \mathbf{1} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}{}^{T} \right] \tag{15}$$

where $\mathbf{1}$ is a matrix of suitable size with all elements equal to 1; $.*$ and $./$ are element-wise multiplication and division,

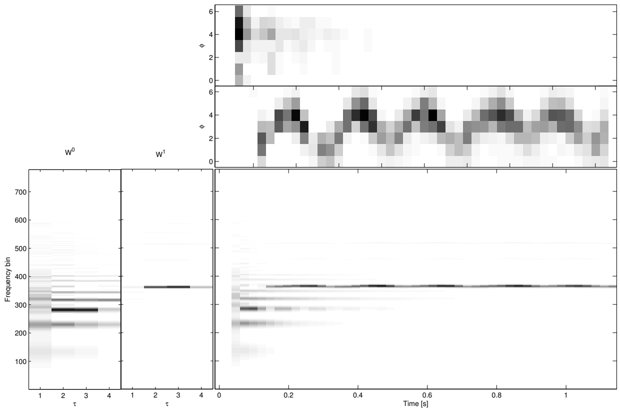

Fig. 2. Factorization of the piece of music using NMF2D. The two time-frequency plots on the left are $\mathbf{W}^{\tau}$ for each factor, i.e. the time-frequency signature of the two sources. The two time-pitch plots on the top are $\mathbf{H}^{\phi}$ for each factor showing how the sources are placed in time and pitch.

It is shown that NMF2D is an effective method for audio separation [2] because it enables to represent a regularly repeating pattern that spans multiple columns(rows) of the spectrogram using multiple bases(gains) that describe the entire sequence. To give an idea, the basis matrices $\mathbf{W}^{\tau}$ and the gain matrices $\mathbf{H}^{\phi}$ for each factor are displayed in Fig. 2 together with the mixture spectrogram. The N columns of $\mathbf{W}^{\tau}$ obtained by NMF2D represent the dominant spectral patterns contained in the input whereas their weights $\mathbf{H}^{\phi}$ correspond to their temporal profiles.

In the proposed method, NMF2D is performed on the auditory spectrogram (sonogram) of the mixture signal for decomposing the sonogram of each component. The sonograms constituting the same audio source are then clustered into a single source sonogram by clustering the base matrices using the K-means clustering algorithm [10]. Each time-domain source signal is reconstructed by masking the mixture spectrogram $\mathbf{X}$ depending on the source sonogram. First, we reconstructed the sonogram of each component by

$$\Lambda_n[b,i] = \sum_{\tau=0}^{T-1}\sum_{\phi=0}^{\Phi-1} W^{\tau}[b-\phi,n]H^{\phi}[n,i-\tau], \qquad (16)$$

for each specific value of n. Then we clustered the component sonograms constituting the same source by applying K-means clustering on the base matrices. Based on the reconstructed source sonograms, we constructed a spectrogram mask for each source, so that the value of each spectrogram bin is assigned to each source in proportion with the sonogram value at that bin. The mapping of the spectrogram masks back into the spectrogram domain is performed as it is proposed in [2]. The complex spectrogram is filtered based on the masks, and the inverse filtered spectrogram is computed using the mixture phase .

3. SIMULATION RESULTS

To test the proposed approach, monophonic mixtures were synthetically generated by summing two different sources. Different datasets have been considered and described below.

- **Dataset A** consists of synthetic mono mixtures of $N = 2$ sources (piano and drums) created using 10 seconds-excerpts of original separated tracks from the song "Sunrise" by S. Hurley, available under a Creative Common License at [11] and downsampled to 16 kHz.
- **Dataset B** consists of synthetic mixtures of speech and music sources obtained from the development dataset of the Signal Separation Evaluation Campaign (SiSEC 2008)[12].
- **Dataset C** consists of synthetic mixtures of two speech sources obtained from the development dataset of the Signal Separation Evaluation Campaign (SiSEC 2008)[12].

All the sources are 10 seconds-long and sampled at 16 kHz.

The separation is performed using the method outlined in Section 2 by applying a pre-processing on the spectrogram. The performance of the proposed method (NMF2D-sone) is compared with the conventional NMF2D model applied on the spectrogram (no pre-processing) in order to see the contribution of the psychoacoustic pre-processing.

In order to evaluate the quality of the separated sources we use the Signal-to-Distortion-Ratio (SDR), Signal-to-Interference-Ratio (SIR) and Signal-to-Artifacts-Ratio (SAR). We used MATLAB routines for computing these criteria obtained from the SISEC'08 webpage [12] and reported the results in terms of SIR, SAR and SDR.

The observation signal is represented using the log-magnitude spectrogram. The audio signals are analyzed by the short time Fourier transform with a $N_F = 2048$ point Hanning windowed FFT and 50% overlap. $N_F/2+1 = 1025$ FFT slices are obtained. The spectrogram bins are grouped into $\lfloor \log \frac{F_S/2}{80} / \log 2^{(1/48)} \rfloor = 318$ logaritmically spaced frequency bins in the range of 80 Hz to 8 kHz with 48 bins per octave, which corresponds to four times the resolution of the equal tempered musical scale. Then, we performed the NMF2D analysis of the log-frequency magnitude spectrogram. For the remaining parameters, we used the following values, rank = [2 20 80 200], τ = [0 1 5 9], ϕ= [0 1 5 9]. The sampling rate of the inputs was 16 kHz. We performed separation using all combinatios of these parameters on our dataset which amounted to 64 experiments for each of three mixtures, repeated 10 times for a total of 1920 experiments. We averaged the performance measures over all the experiments and analyzed the effect of various parametes.

Of more importance than the individual parameters is the interaction between them. Fig. 3-5 are instrumental in pointing this out. We briefly describe some of the major interactions here. We also reported the results of the conventional NMF2D algorithm proposed in [2] (NMF2D) for comparison issues. As it is seen from the figures, the length of the bases (τ) varies the performance measures significantly. The proposed NMF2D-sone algorithm performs better for $\tau = \{0,1\}$. Similarly, the NMF2D method also performs better for low τ values. If we compare the results depending on the number of gains ϕ, we see that the proposed method performs better SDR and SIR values at $\phi = \{0,1\}$. The rank parameter is also of main importance. The proposed method performs higher SDR and SIR values for rank $r = 2$. If we compare the performance of the proposed method to the conventional NMF2D algorithm, we can see that the proposed method increases the SDR value by an amount of 0.7 dB. The dependency of the NMF2D performance on the same parameters is quite different. NMF2D method performs better for low τ and high ϕ values. The performance of NMF2D increases as rank increases. The highest SDR, SIR and SAR values are obtained at around rank $r = 20$. The performance starts to decrease slightly if we increase the rank much more.

Figure 4 illustrates the performance of the proposed method in terms of SIR. We can see the same characteristics depending on the parameters of τ, ϕ and rank. However, the SIR values obtained by NMF2D method is slightly better than the SIR values obtained by the proposed method.

The change in SAR values for the parameters τ, ϕ and rank are completely different than SDR and SIR. The proposed method performs higher SAR values for $\tau = \{0,1\}$, $\phi = \{5,9\}$ and rank $r = \{80,200\}$. The increase in SAR values obtained by the proposed method is huge and it is around 7 dB.

In general, as τ grows, more bases acted as a detriment to the separation quality. More gains introduced less interference noise, which causes SDR and SIR to increase and SAR to decrease. Finally as the rank increases we noted that the number of bases and number of gains become more important.

In order to measure the computational load of the proposed method, we implemented the NMF2D and NMF2D-sone algorithms for a particular experiment using 10 sec mono mixture (sampled at 16 kHz) with $N = 2$ sources and NMF2D parameters selected as $r = 200$ components, $\tau = 0$ and $\phi = 9$. The algorithms are run on a PC equipped with a 2.4 GHz Intel Core2 Quad processor using the same parameters and initial conditions.

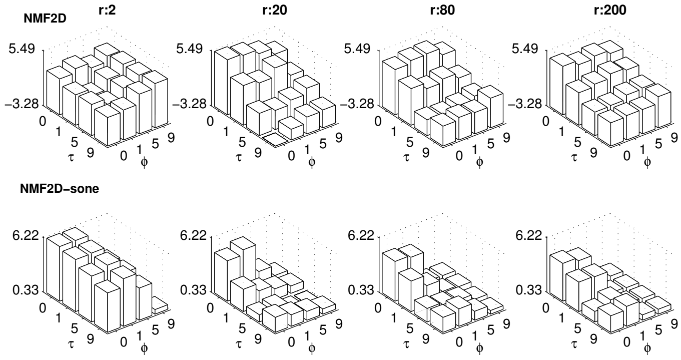

Fig. 3. Performance measures in terms of SDR for combinations of NMF2D parameters τ and ϕ. Each row of plots is for a different method as denoted over each row (NMF2D, NMF2D-sone). Each column is obtained by a different rank number as denoted over each column ($r = 2, r = 20, r = 80, r = 200$). Note that all plots are plotted across different scales indicated by their maximum and minimum values along the vertical axis.

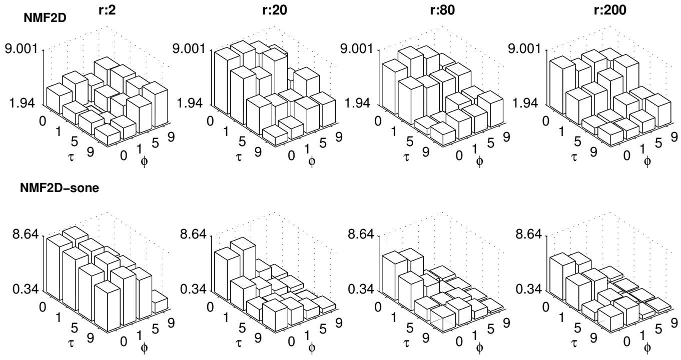

Fig. 4. Performance measures in terms of SIR for combinations of NMF2D parameters τ and ϕ. Each row of plots is for a different method as denoted over each row (NMF2D, NMF2D-sone). Each column is obtained by a different rank number as denoted over each column ($r = 2, r = 20, r = 80, r = 200$). Note that all plots are plotted across different scales indicated by their maximum and minimum values along the vertical axis.

NMF2D takes about 1879.24 sec to converge in 2000 iterations for this particular experiment. The proposed NMF2D-sono algorithm takes about 48.59 sec to converge in 185 iterations. NMF2D-sono algorithm decreases the size of the data matrix, thus decreases the computational time while increasing the separation quality of the estimated sources significantly.

4. CONCLUSION

In this paper, we propose a perceptual audio source separation method using NMF2D. The perceptuality is integrated into the separation algorithm by psychoacoustic pre-processing applied on the mixture spectrogram. The simulation results indicate that the proposed psychoacoustic pre-processing significantly improves the quality of the reconstructed audio sources and decreases the computational complexity.

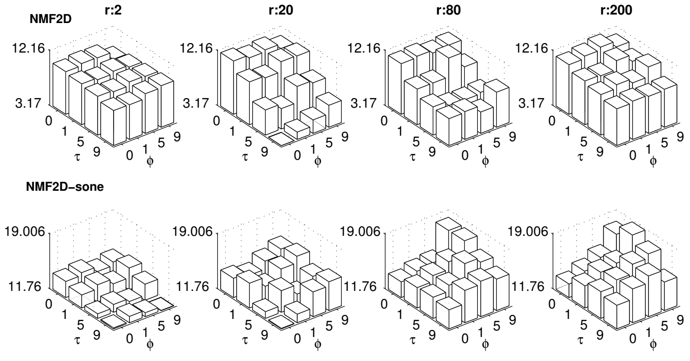

Fig. 5. Performance measures in terms of SAR for combinations of NMF2D parameters τ and ϕ. Each row of plots is for a different method as denoted over each row (NMF2D, NMF2D-sone). Each column is obtained by a different rank number as denoted over each column ($r = 2, r = 20, r = 80, r = 200$). Note that all plots are plotted across different scales indicated by their maximum and minimum values along the vertical axis.

5. REFERENCES

[1] T. O. Virtanen, "Monaural sound source separation by perceptually weighted non-negative matrix factorization," Tech. Rep., Tampere University of Technology, 2007.

[2] M.N.Schmidt and M.Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. of ICA'06*, Charleston, SC, USA, 2006.

[3] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *LNCS*. 2004, pp. 494–499, Springer-Verlag Berlin Heidelberg.

[4] T. O. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, vol. 22, Springer Series of Information Sciences, 1999.

[6] R.R. Guddeti and B. Mulgrew, "Perceptually motivated blind source separation of convolutive mixtures," in *Proc. of ICASSP2005*, Philadelphia, PA, USA, 2005.

[7] E. Pampalk, "A matlab toolbox to compute music similarity from audio," in *Proc. of ISMIR'03*, Baltimore, MD, 2003, pp. 201–208.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing*, vol. 13, 2001.

[9] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, pp. 155–182, 1979.

[10] J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, Inc., New York, 1975.

[11] S. Hurley, "Call for remixes: Shannon hurley," Available: http://ccmixter.org/shannon-hurley.

[12] "Signal separation evaluation campaign (SISEC 2008)," Available: http://sisec.wiki.irisa.fr, 2008.