AN ADAPTIVE TIME-FREQUENCY RESOLUTION FRAMEWORK FOR SINGLE CHANNEL SOURCE SEPARATION BASED ON NON-NEGATIVE TENSOR FACTORIZATION

S. Kırbız, B. Günsel

Multimedia Signal Processing and Pattern Recognition Group.
Dept. of Electronics and Comm. Eng. İstanbul Technical University, Turkey

ABSTRACT

In this paper, we propose an adaptive time-frequency resolution based single channel sound source separation method using Non-negative Tensor Factorization (NTF). The model aims to alleviate drawbacks of working by fixed length Short Time Fourier Transform (STFT) by minimizing the smearing of signal energy in both time and frequency. A joint optimization scheme has been applied based on KL-divergence where each layer of the tensor represents the mixture at a different resolution. In order to enclose sparseness into factorization, the resynthesis is made through an adaptive weighted fusion procedure which combines the separated sources in a manner that maximizes the energy concentration. Test results reported over a large sound database indicate the introduced NTF based fusion method improves the sound quality both in terms of conventional and perceptual distortion measures.

Index Terms— Non-negative Tensor Factorization, sound source separation, adaptive time-frequency resolution

1. INTRODUCTION

Existing audio source separation methods mostly use a time-frequency representation of the signal such as spectrogram derived by the Short-Time Fourier Transform (STFT). The main problem with the fixed time-frequency resolution STFT is the smearing of signal energy in either direction. Smearing in time causes artifacts such as pre- or post-echoes around transients which lead to incorrect detection of temporal changes [1]. On the other hand, smearing of signal energy in frequency prevents distinguishing closely spaced harmonics.

This paper introduces a method that aims to alleviate problems encountered in single channel sound source separation techniques. We propose an adaptive short-time analysis-synthesis scheme in order to arrive at a signal-dependent supervised source separation method which reduces the artifacts caused by using a single resolution based representation. In our approach, the multiresolution time-frequency representation of the observed signal is represented as an "n-way array" or in other terms as a "tensor", where each layer of the tensor denotes the magnitude spectrogram of the observed mixture obtained using a different time-frequency resolution. Since the input is represented as a multidimensional matrix, Non-negative Tensor Factorization (NTF) [2] which is a natural generalization of Non-negative Matrix Factorization (NMF) in higher dimensional spaces is preferred in content representation. NTF has been widely used to separate multichannel recordings [3] using the information from different observations by joint optimization. Different from the separation methods based on NTF, the proposed scheme uses only a single observation represented at various time-frequency resolutions and enhances the quality of the separated sources compared to the NMF based methods which use the information from a single observation. The convergence of the NTF algorithm yields the separated sources in each time-frequency resolution. After reconstructing the sources in various resolutions, the adaptation is performed based on a measure of energy concentration as it is performed in [4].

There are some works which deal with the limitations of fixed time-frequency resolution in source separation. In [5], Wavelet Packet (WP) transform which is a multiscale transform is used to decompose signals into sets of local features with various degrees of sparsity. Then, the best subset is selected from an overcomplete set of WP features of mixtures with respect to the estimation of the separation error and used for separation. In our previous work [4], the separated sources obtained by applying NMF at different time-frequency resolutions are adaptively fused based on maximum energy compaction principle. The method proposed in this work combines the parallel NMF factorizations introduced in [4] into a single NTF scheme, thus performs a joint optimization by fusing the information from various time-frequency resolutions. It is shown that the proposed scheme enhances the quality of the separated sources.

2. PROPOSED METHOD

We introduce a MultiResolution NTF (MR-NTF) method which performs separation of learned sources from monophonic mixtures based on adaptive time-frequency resolution. The proposed MR-NTF approach aims to optimize a generalized Kullback-Leibler (KL) divergence:

$$D = \sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(X_{cki}\log\frac{X_{cki}}{\hat{X}_{cki}} - X_{cki} + \hat{X}_{cki}\right), \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{C\times K\times I}$ is the observed mixture represented at C time-frequency resolutions having K frequency bins and I time frames. $\hat{X}_{cki} = \sum_{r=1}^{R} Q_{cr}S_{kr}A_{ir}$ is an approximation to X_{cki}, where $\mathbf{Q} \in \mathbb{R}^{C\times R}$ is a matrix containing the gains of each component in each layer, $\mathbf{S} \in \mathbb{R}^{K\times R}$ is the basis matrix containing the frequency basis vectors and $\mathbf{A} \in \mathbb{R}^{I\times R}$ is the corresponding amplitude envelopes for the frequency basis vectors and R is the rank of factorization.

The proposed method learns codebooks $\mathbf{S}$ of sources from time-frequency representations of their training signals via NMF at various resolutions. Thus, the KL divergence is optimized using multiplicative update rules for the Q_{cr} and A_{ir}:

$$Q_{cr} \leftarrow Q_{cr} * \left(\sum_{k=1}^{K} S_{kr}\sum_{i=1}^{I}\Lambda_{cki}A_{ir}\right)\Big/\left(\sum_{k=1}^{K} S_{kr}\sum_{i=1}^{I}A_{ir}\right)$$

$$A_{ir} \leftarrow A_{ir} * \left(\sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}\Lambda_{cki}S_{kr}\right)\Big/\left(\sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}S_{kr}\right),$$

where $\Lambda_{cki} = X_{cki}/\hat{X}_{cki}$. The $\mathbf{Q}, \mathbf{S}$ and $\mathbf{A}$ factors are then applied on magnitude spectrogram tensor $\mathbf{X}$ of the mixture to extract source estimates. Finally, we fuse the output from multiple resolutions to obtain more robust estimates of the sources using the maximal energy compaction principle method used in [1], [4] for varying the time-frequency resolution adaptively. This approach estimates the sparsity of different time-frequency resolutions and mixes them accordingly so as to obtain minimal smearing both in time and frequency directions.

2.1. Fusing the Time-Frequency Resolutions by MR-NTF

In particular, the MR-NTF bases of the $j-$th source ($j = 1\cdots J$) are learned from corresponding magnitude spectrograms $\mathbf{X}_{jc}$ at various resolutions $c = 1\cdots C$ using NMF in the Dictionary Learning block of Fig.1. Bases extracted at different resolutions are combined into a matrix $\mathbf{S} = [\mathbf{S}_1\cdots\mathbf{S}_{JR}]$ which is of size $K \times JR$. Note that, the observed mixture signal $\mathbf{x}(t)$ does not include the source signals used for learning. The MR-NTF is applied on magnitude spectrogram tensor $\mathbf{X}$, which is constructed by concatenating the spectrograms of observed mixture at various resolutions, by fixing the bases to $\mathbf{S} \in \mathbb{R}^{K\times JR}$ and learning their amplitudes $\mathbf{A} \in \mathbb{R}^{I\times JR}$ and the gains $\mathbf{Q} \in \mathbb{R}^{C\times JR}$ of each factor in each resolution.

Fig.1 illustrates main steps of the NTF-based source separation algorithm running on the multiresolution representation of the input mixture. In the figure, only two resolutions are depicted for clarity, but the framework can be extended to any number. The algorithm proposed for automatically separating the sources is as follows:

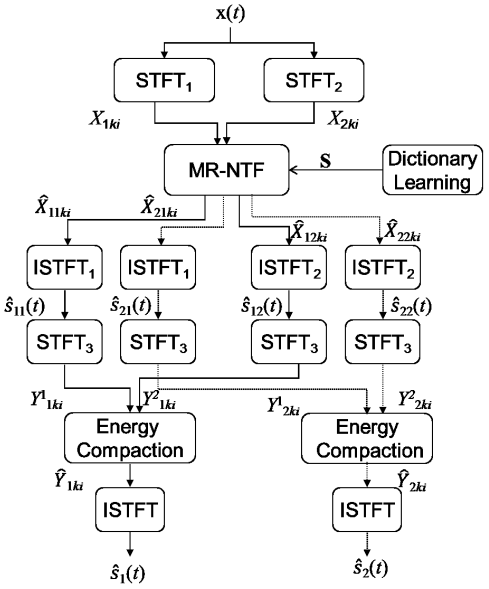

Fig. 1. General scheme for the proposed MR-NTF.

1. The spectrograms $\mathbf{X}_c, c = 1\cdots C$ at different time-frequency resolutions are obtained. The hop size and the frequency grids should be equal for all C STFT resolutions in order to ensure that all STFT magnitudes $\mathbf{X}_c$ are calculated in the same grid of time-frequency locations. In order to achieve that, we also zero pad the smaller STFT windows to ensure that all of the STFTs will have the same number of frequencies. The $(k,i)-$th time-frequency component of the mixture spectrograms obtained at two different resolutions are represented as X_{cki} where $c = \{1, 2\}$ in Fig.1.

2. The spectrograms are combined into a tensor $\mathbf{X}$, where the $c-$th layer of $\mathbf{X}$ represents the mixture spectrogram $\mathbf{X}_c$ calculated based on the $c-$th resolution.

3. The bases learned at the dictionary learning block are fixed as $\mathbf{S} = [\mathbf{S}_1\cdots\mathbf{S}_{JR}]$ where each column of $\mathbf{S}$ has K frequency components.

4. MR-NTF is performed on the tensor $\mathbf{X} \in \mathbb{R}^{C\times K\times I}$ by fixing $\mathbf{S} \in \mathbb{R}^{K\times JR}$ and updating $\mathbf{A} \in \mathbb{R}^{I\times JR}$ and $\mathbf{Q} \in \mathbb{R}^{C\times JR}$ at each iteration.

5. Upon convergence, $\mathbf{A} = [\mathbf{A}_1\cdots\mathbf{A}_{JR}]$ and $\mathbf{Q} = [\mathbf{Q}_1\cdots\mathbf{Q}_{JR}]$ are segmented into sources, each $\mathbf{A}_j, \mathbf{Q}_j$, $j = 1\cdots J$ corresponding to one source.

6. The contribution of each source in the mixture magnitude spectrogram at the $c-$th resolution are estimated from:

$$\hat{X}_{jcki} = \sum_{r=1}^{R} Q_{jcr}S_{jkr}A_{jir} \quad (2)$$

where S_{jkr} represents the $k-$th frequency component of the $r-$th basis vector of the $j-$th source; Q_{jcr} represents the gain of the $r-$th component of the $j-$th source in the $c-$th resolution; A_{jir} represents the time envelope for the $r-$th component of the $j-$th source in the $i-$th time frame.

7. Complex source spectrograms are directly retrieved using Wiener filter [6]

$$\hat{F}_{jcki} = F_{cki}\frac{\hat{X}_{jcki}}{\sum_j \hat{X}_{jcki}}, \quad (3)$$

where F_{cki} represents the $k-$th frequency component of $i-$th time frame for the mixture signal represented at the $c-$th resolution.

8. Time-domain estimate $\hat{s}_{jc}(t)$ of the $j-$th source at the $c-$th resolution is obtained by applying Inverse STFT (ISTFT) on estimated coefficients $\hat{F}_{jcki}$.

2.2. Enhanced Sparsity by Maximal Energy Compaction

Our goal is to combine the source signals from different resolutions in order to achieve a compact sparse representation in each part of the time-frequency plane. This combination is performed using the maximal energy compaction principle as in [1] by an additional filter bank, with a fixed time-frequency resolution that transforms the resulting separated source signals into time-frequency coefficients on the same time-frequency grid as in the analysis steps. This is achieved by first transforming the estimated time-domain sources $\hat{s}_{jc}(t)$ into the time-frequency domain by using a fixed time-frequency resolution. The resulting STFTs $Y^c_{jki}, j = 1 \cdots J$ correspond to the time-frequency representations of the sources obtained from MR-NTF in the $c-$th layer.

In order to fuse the information efficiently at every time-frequency bin (k, i) we consider a rectangular area Ω around this point. There is a trade-off between selecting a small or a large area. If the area is small, there won't be enough coefficients to calculate a robust estimate of energy smearing. If it is too big, it will not be a local estimate. Less smearing around a time-frequency component yields a sparser representation, thus maximizes the energy compaction. In order to estimate the sparsity in a rectangular grid $\Omega = H \times U$, we use a method based on kurtosis [1],[4]:

$$K^c_{jki} = \frac{\frac{1}{HU}\sum_{k',i' \in \Omega}(|Y^c_{jk'i'}|^2 - |\bar{Y}^c_j|)^4}{\left(\frac{1}{HU}\sum_{k',i' \in \Omega}(|Y^c_{jk'i'}|^2 - |\bar{Y}^c_j|)^2\right)^2}, \quad (4)$$

where $|\bar{Y}^c_j|$ is the sample mean of squared STFT magnitudes $|Y^c_{jki}|^2$ in the grid Ω. Kurtosis is widely used for measuring the nongaussianity of a distribution and it grows with sparsity resulting in peaky distributions [7].

In order to avoid hard switching from one resolution to another, we fuse the squared magnitude coefficients from different resolutions. The fusion is performed by a weighted sum of the squared magnitude spectrogram coefficients:

$$|Y_{jki}|^2 = \sum_{c=1}^{C} w^c_{jki}|Y^c_{jki}|^2, \quad (5)$$

where Y^c_{jki} is the $k-$th frequency component at the $i-$th time frame of the $j-$th source obtained for the $c-$th fixed time-frequency resolution and the mixing weights are calculated as:

$$w^c_{jki} = \frac{K^c_{jki}}{\sum_{c=1}^{C} K^c_{jki}}. \quad (6)$$

This step yields a single power-magnitude spectrogram $|Y_{jki}|^2$ for each source. In (5), the estimated sources from different time-frequency resolutions are combined in an adaptive way, such that the smearing in both time and frequency is minimized.

The adaptive representation requires the phase information in order to estimate the time-domain sources. Wiener filtering of the mixture spectrogram $\mathbf{Y}$ is performed such as

$$\hat{Y}_{jki} = Y_{ki}\frac{|Y_{jki}|^2}{\sum_{j=1}^{J}|Y_{jki}|^2}. \quad (7)$$

The ISTFT is then applied on the complex STFT coefficients $\hat{Y}_{jki}$ in order to transform the extracted sources $\hat{Y}_{jki}$ to time domain signals $\hat{s}_j(t)$.

3. PERFORMANCE EVALUATION

In order to test the proposed approach, ten monophonic mixtures are synthetically generated by summing $J = 2$ different but equal length sentences uttered by male and female speakers from the TIMIT database. A training data of length 21 to 33 sec is used for each speaker in order to learn the bases of each speaker. The length of the evaluation sentences are 2 to 3 sec long. All the audio files are sampled at 16 kHz. Note that, the evaluation data is not included in the training set. Evaluation of the quality of speech separation algorithms is performed using Signal-to-Distortion-Ratio (SDR), Signal-to-Interference-Ratio (SIR) and Signal-to-Artifacts-Ratio (SAR) [8] and their perceptual correspondences Overall Perceptual Score (OPS), Interference-related Perceptual Score (IPS) and Artifacts-related Perceptual Score (APS) [9].

The separation is performed using the proposed method described in Section 2. The data is analyzed using a Hanning windowed STFT of $N_F = \{512, 2048\}$ samples. The effect of the rank of the factorization is investigated by selecting the training number of bases for each source as $R = \{1, 10, 20, 50, 100\}$. Another important parameter is related to the adaptive mixing. The adaptive mixing is performed based on a sparsity measure calculated over a time-frequency

grid Ω around each time-frequency component with various sizes. As a compromise between sparseness and computational complexity, $P = 3$ and $T = 3$ are selected for the grid width (frequency components) and height (time frames) to calculate the kurtosis around each time-frequency component.

We run the proposed MR-NTF algorithm for 50 iterations. We perform separation using all combinations of the parameters on our dataset which amounted to 45 experiments (five rank values, nine different grid size) for 10 test signals using two different resolutions represented by $C = 2$ layered tensors. We report the mean of the performance measures for all these experiments on our dataset. In order to investigate the improvement in the quality of the estimated sources obtained by the proposed approach, we compare the results with NMF results obtained at a single resolution and our previous MR-NMF work [4]. In MR-NMF, different from the MR-NTF, the multiple NMF instances are run independently and the sources obtained from different resolutions are merged as it is described in Section 2.2

In Fig.2, we investigate the effect of the rank parameter on separation performance. Thus, $R = \{1, 10, 20, 50, 100\}$ bases are learned for each source in the training and used for separating the sources. The horizontal axes in Fig.2 display the number of bases per each source. In the figure, the results obtained by fixed time-frequency resolutions by NMF are also plotted to show the improvement achieved by mixing the fixed time-frequency resolution results in an adaptive way. The best performance is obtained when $C = 2$ where $N_F = 512$ and $N_F = 2048$ are used for representing the mixture signal in different layers of the input tensor. Thus, only the results obtained on a $C = 2$ layer tensor where $N_F = \{512, 2048\}$ are used to represent the input mixture are reported. As it is seen from Fig.2, the proposed method increases the separation quality by around 2 dB and 3 dB in terms of SDR and SIR while the SAR is decreased. If we compare MR-NMF and MR-NTF results, we can see that both methods outperform the results obtained from a fixed resolution. We also observe that MR-NTF outperforms MR-NMF by 1-2 dB in terms of SDR and SIR for rank values of 10 and 20. We can also conclude that, learning $R = 20$ bases for each source gives the optimum results in terms of all measures which also gives us the opportunity to separate the sources with a lower computational complexity. In Table 1, we report the quality of the separated sources in terms of mean SDR, SIR and SAR and their perceptual correspondences OPS, IPS and APS obtained from one speech mixture where rank is selected as $R = 20$ for each source. The results obtained by NMF for two different resolutions are reported in the second and third rows of the table in terms of SDR, SIR, SAR, OPS, IPS and APS. The adaptive results obtained by our MR-NTF method are reported on the fourth row. We observe that, the proposed adaptive MR-NTF scheme improves the quality of the separated sources by 1-2 dB in terms of SDR and 4-5 dB in terms of SIR compared to fixed-resolution results. The artifacts obtained for the adaptive resolution scheme and the fixed-resolution are similar. It is also observed that, the OPS and IPS values obtained by MR-NTF and MR-NMF methods are similar to the highest OPS and IPS values obtained by NMF at different time-frequency resolutions. The artifact is increased in the proposed approach by an amount of 1-24 APS.

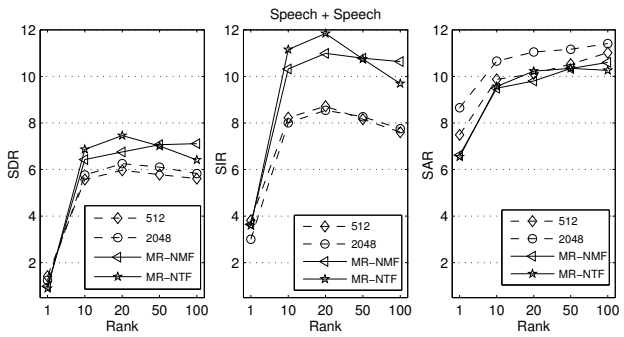

Fig. 2. The mean SDR, SIR and SAR values obtained for various number of bases per each source obtained using a fixed time-frequency resolution NMF, MR-NMF and MR-NTF.

Table 1. Performance in terms of SDR, SIR, SAR, OPS, IPS and APS values obtained by NMF, MR-NMF [4] and MR-NTF methods on a mixture of female and male speaker.

	SDR	SIR	SAR	OPS	IPS	APS
NMF(512)	8.2	12.0	10.9	**45.4**	61.1	44.1
NMF(2048)	8.7	11.9	**11.8**	29.3	39.9	**67.8**
MR-NMF	9.6	16.1	11.0	44.5	**65.1**	42.4
MR-NTF	**10.0**	**16.6**	11.4	45.0	60.0	43.5

4. CONCLUSION

In this paper, we present an adaptive time-frequency resolution supervised method for separating known types of sounds from a single observation. We represent a single mixture signal in various time-frequency resolutions in different layers of the tensor. Then, we perform the proposed MR-NTF approach in order to extract the sources. We observe an improvement of 1-4 dB in terms of SDR and SIR relative to the fixed time-frequency resolution separation results obtained by NMF. In terms of perceptual measures OPS and IPS, the improvement is around 5-30. We also evaluate the performance of the proposed approach on musical signals and observe a similar improvement in the separation quality. However, since the space is limited, we only report the results obtained on a speech database.

5. REFERENCES

[1] A. Lukin and J. Todd, "Adaptive time-frequency resolution for analysis and processing of audio," 120th Audio Engineering Society Convention, Paris, France, May 2006.

[2] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[3] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. of ISSC*, Dublin, Sept. 1-2 2005.

[4] S. Kirbiz and P. Smaragdis, "An adaptive time-frequency resolution approach for non-negative matrix factorization based single channel source separation," in *Proc. of ICASSP*, May 22-27 2011, pp. 253–256.

[5] P. Kisilev, M. Zibulevsky, and Y. Y. Zeevi, "Multiscale framework for blind source separation," *The Journal of Machine Learning Research*, vol. 4, no. 7-8, pp. 1339–1364, 2004.

[6] D. FitzGerald and M. Cranitch, "Resynthesis methods for sound source separation using shifted non-negative factorisation models," in *Proc. of ISSC*, Derry, Sept 13-14 2007.

[7] C. C. Took and S. Sanei, "Exploiting sparsity, sparseness and super-gaussianity in underdetermined blind identification of temporomandibular joint sounds," *Journal of Computers*, vol. 2, no. 6, pp. 65–71, August 2007.

[8] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[9] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Jan. 2011.