AN ADAPTIVE TIME-FREQUENCY RESOLUTION APPROACH FOR NON-NEGATIVE MATRIX FACTORIZATION BASED SINGLE CHANNEL SOUND SOURCE SEPARATION

Serap Kırbız[*]

İstanbul Technical University, Turkey

Paris Smaragdis

University of Illinois at Urbana-Champaign
& Adobe Systems Inc.

ABSTRACT

In this paper, we propose an adaptive time-frequency resolution approach for the single channel source separation problem. The aim is to improve the quality and intelligibility of the separated sources by adapting the time-frequency resolution of the analysis window to the characteristic of the signal under consideration. The results evaluated on a large test set show the improvements obtained by the proposed algorithm.

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

1. INTRODUCTION

Audio source separation has been used in many applications such as structured audio coding, automatic transcription of music and robust speech recognition.

In this paper we focus on the separation of two speakers from a single monophonic recording. Even though a human listener can easily attend to one speaker or the other, the design of a computational system that has that ability is still an open problem in audio research.

A vast amount of research has been conducted both in blind [1] and supervised [2] single channel audio source separation. Among these, Non-negative Matrix Factorization (NMF) has been widely preferred for its simplicity and efficiency to factorize the mixture signal into a linear combination of basis vectors under non-negativity constraints of output matrices [1], [2]. Usually, NMF is applied on the fixed resolution time-frequency representation of the mixture signal. However, it is well known that vowels can be stationary over long segments, thus employing longer windows leads to improved frequency resolution. On the other hand, the transients smear the time resolution and require to be analyzed in shorter windows [3]. Therefore, it is desirable to change the time-frequency resolution of the mixture signal adaptively based on the signal characteristics.

In this paper, we propose an adaptive time-frequency resolution based single channel supervised source separation scheme using NMF in order to enhance the separation quality. The learning approach benefits from knowledge extracted from each speaker's utterances outside of the input samples based on the method proposed in [2]. The adaptation is based on the maximal energy compaction principle method proposed in [3]. We run multiple instances of NMF algorithm on the same mixture signal with different time-frequency resolutions. We mix the separated source signals obtained from different time-frequency resolutions to obtain minimal smearing both in time and frequency directions. We show that the proposed method leads to an increase of 1-3 dB in Signal-to-Distortion-Ratio (SDR) and Signal-to-Interference Ratio (SIR) relative to the fixed time-frequency resolution approach.

2. SOURCE SEPARATION USING NON-NEGATIVE MATRIX FACTORIZATION

In this section, we describe the non-negative matrix factorization (NMF), which we will employ as the core of our source separation algorithm.

NMF based source separation algorithms aim to factorize an observed magnitude spectrogram matrix $\mathbf{X} \in \mathbb{R}^{F \times T}$ as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{F \times R}$ and $\mathbf{H} \in \mathbb{R}^{R \times T}$ such that

$$X[f,t] \approx \sum_{r=0}^{R-1} W[f,r]H[r,t], \tag{1}$$

where $W[f,r]$ denotes the $f-$th frequency of the $r-$th audio component, $H[r,t]$ is the gain of the $r-$th component in $t-$th time frame and R is the number of audio components. Using prior information, as shown in section 4.1, audio components belonging to the same source are clustered into a single group and they collectively define that source in the mixture. The optimization objective is to minimize the reconstruction error as measured by a Kullback-Leibler-like divergence measure which is defined as:

$$D_{KL} = \sum_{f,t} \left(X_{ft} \log \frac{X_{ft}}{(WH)_{ft}} - X_{ft} + (WH)_{ft} \right). \tag{2}$$

[*] This work is supported by The Council of Higher Education of Turkey.

This function is optimized using multiplicative update rules for the two factors $\mathbf{W}$ and $\mathbf{H}$:

$$\begin{array}{l} H_{rt} \leftarrow H_{rt} \frac{\sum_f W_{fr} X_{ft}/(WH)_{ft}}{\sum_f W_{fr}} \\ W_{fr} \leftarrow W_{fr} \frac{\sum_t H_{rt} X_{ft}/(WH)_{ft}}{\sum_t H_{rt}} \end{array}. \quad (3)$$

Both of these updates are applied iteratively until the two factors converge [4].

3. ADAPTIVE TIME-FREQUENCY RESOLUTION

The main problem with the fixed time-frequency resolution Short Time Fourier Transform (STFT) is the smearing of signal energy. Smearing in frequency can prevent distinguishing closely spaced harmonics and smearing in time can affect estimation of positions and durations of transients. When used in the context of source separation, this smearing prohibits accurate editing in the time-frequency space and can impede performance. Our objective in this paper is to alleviate that problem by estimating the smearing amount for different fixed time-frequency resolutions and selecting the resolution that minimizes smearing in both time and frequency.

We use the maximal energy compaction principle method proposed in [3] 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.

3.1. Maximal Energy Compaction Principle

In this section, we describe the maximal energy compaction principle proposed in [3]. First, we calculate the fixed time-frequency resolution STFT coefficients for different resolutions using time-domain windows of varying length. The hop size and the frequency grids should be equal for all L STFT resolutions in order to ensure that all STFT squared magnitudes $|\mathbf{X}_l|^2, l = 1 \cdots L$ 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.

In order to estimate the sparsity in a rectangular grid $\Omega = P \times Q$ around a specified time-frequency bin (f, t), we used three different methods which are compared in [5]. The first method computes the L_2 norm over L_1 norm of squared STFT magnitude $|X_l(f,t)|^2$ in the grid Ω as:

$$S_l^N[f,t] = \frac{\sqrt{\sum_{f',t' \in \Omega} |X_l[f',t']|^4}}{\sum_{f',t' \in \Omega} |X_l[f',t']|^2}, \quad (4)$$

where l denotes a fixed time-frequency resolution. The second method is based on kurtosis:

$$S_l^K[f,t] = \frac{\frac{1}{PQ}\sum_{f',t' \in \Omega}(|X_l[f',t']|^2 - \bar{X}_l)^4}{\left(\frac{1}{PQ}\sum_{f',t' \in \Omega}(|X_l[f',t']|^2 - \bar{X}_l)^2\right)^2}, \quad (5)$$

where $\bar{X}_l$ is the sample mean of squared STFT magnitudes $|X_l(f,t)|^2$ in the grid Ω. The third method is based on entropy:

$$S_l^E = - \sum_{f',t' \in \Omega} (|X_l[f',t']|^2 \log |X_l[f',t']|^2). \quad (6)$$

All the methods mentioned above are widely used as sparsity measures in different applications [5].

3.2. Resolution Selection Algorithm

In order to obtain the optimal resolution at every time-frequency bin (f, t) 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.

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

$$|X[f,t]|^2 = \sum_{l=1}^{L} w_l[f,t] |X_l[f,t]|^2, \quad (7)$$

where $|\mathbf{X}_l|^2$ is the squared STFT magnitude for the $l-$th fixed time-frequency resolution and the mixing weights are calculated as:

$$w_l[f,t] = \frac{S_l[f,t]}{\sum_{l=1}^{L} S_l[f,t]}. \quad (8)$$

$S_l[f,t]$ is a measure of sparsity calculated using one of the three methods explained in Section 3.1.

4. PROPOSED METHOD

In this section, we will introduce a way to incorporate adaptive time-frequency resolution when performing separation from monophonic mixtures of known speakers. This process involves learning NMF codebooks of known speakers at different time-frequency resolutions, applying these on corresponding spectrograms of a mixture to extract speaker estimates, and finally adaptively interpolating the output from the multiple resolutions to obtain more robust estimates of the sources.

4.1. Separation of Known Speakers

In [2], it is shown that once the basis functions of each speaker are known they can be used to reconstruct the speaker's signal from a monophonic mixture. Based on the work [2], the bases learned separately from the male ($\mathbf{W}_m$) and the female ($\mathbf{W}_e$) speakers are used to reconstruct the source signal from their mixture. This can be performed by learning each speaker's NMF bases from matrices containing their magnitude spectrograms $\mathbf{X}_m$ and $\mathbf{X}_e$ using the method shown in Section 2. By combining these bases we obtain a union of the bases $\mathbf{W} = [\mathbf{W}_m\mathbf{W}_e]$ which is of size $M \times 2R$. We can then take a mixture of the known speakers uttering unknown phrases $\mathbf{x}(t)$, and perform NMF on its spectrogram $\mathbf{X}$ by fixing the bases to $\mathbf{W}$ and learning their weights $\mathbf{H}$. Upon convergence we segment $\mathbf{H}$ in two parts $\mathbf{H} = [\mathbf{H}_m\mathbf{H}_e]^\top$, each corresponding to one speaker. We can then estimate the contribution of each speaker in the mixture magnitude spectrogram $\mathbf{X}$ using only the individual speaker bases and weights by: $\hat{\mathbf{X}}_m = \mathbf{W}_m\mathbf{H}_m$ and $\hat{\mathbf{X}}_e = \mathbf{W}_e\mathbf{H}_e$.

4.2. Mixing the Single Resolution Results

Building filter banks with variable time-frequency resolution has been commonly addressed for audio compression methods [6]. However, these approaches are limited by the fact that compression requires to keep the size of the data representing the signal at a minimum. On the other hand, audio processing methods such as source separation can benefit from redundancy which leads to multi-resolution framework presented here. In Fig. 1, the same NMF-based source separation algorithm is running with several parallel instances on different fixed time-frequency resolutions of the same input mixture signal. In the figure, only two resolutions are depicted for clarity, but the framework can be extended to any number. First, STFTs with different time-frequency resolutions ($STFT_1$ and $STFT_2$) are applied on the mixture signal in order to get the mixture spectrograms of different time-frequency resolutions ($X_1(f,t)$ and $X_2(f,t)$). NMF-based separation is applied on the resulting mixture spectrograms $X_1(f,t)$ and $X_2(f,t)$ independently in order to obtain estimated source spectrograms as shown in the previous section. The source spectrograms $\mathbf{X}_{m,1}$ ($\mathbf{X}_{m,2}$) and $\mathbf{X}_{e,1}$ ($\mathbf{X}_{e,2}$) represent the estimated source spectrograms of male and female speakers obtained by applying NMF to $\mathbf{X}_1$ ($\mathbf{X}_2$), respectively. An inverse STFT (ISTFT) is then performed on each source spectrogram in order to get the time-domain estimates of the source signals. Note that $\tilde{s}_{m,1}(t)$ ($\tilde{s}_{e,1}(t)$) and $\tilde{s}_{m,2}(t)$ ($\tilde{s}_{e,2}(t)$) correspond to the time-domain estimates of the male (female) speaker using different time-frequency resolutions. Our goal is to combine the source signals from different resolutions in order to achieve an optimally compact representation in each part of the time-frequency plane. This combination is performed as in [3] by an additional filter bank, with a fixed time-frequency resolution that transforms these resulting signals into time-frequency coefficients on the same time-frequency grid as in the analysis steps. The mixing is performed based on the analysis of sparsity of the signal in a time-frequency area as described in Section 3.1. The ISTFT is then applied on the mixed STFT coefficients in order to estimate the source signals.

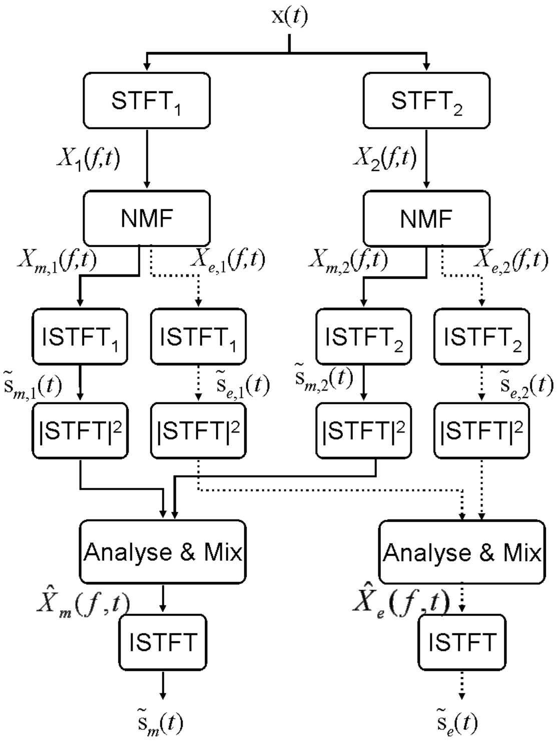

Fig. 1. General scheme for source separation using adaptive multiresolution NMF.

4.3. Resynthesis of the Source Signals

The time-domain estimates of the sources ($\hat{s}_m(t)$ and $\hat{s}_e(t)$) are obtained by calculating masks for each source by comparing the estimated spectrograms ($\hat{\mathbf{X}}_m$ and $\hat{\mathbf{X}}_e$), so that each spectrogram bin is assigned to each speaker proportional to the power of the estimated speaker spectrogram. We then apply a spectral filter on the mixture spectrogram $\mathbf{X}$ based on these masks and compute the inverse filtered spectrograms using the phase of the mixture [1].

5. SIMULATION RESULTS

To test the proposed method, monophonic mixtures are synthetically generated by summing two different but equal length sentences uttered by male and female speakers from the TIMIT database. A training data of length 10 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. 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). We used MATLAB routines for computing these criteria obtained from the SISEC'08 web page [7] and reported the results in terms of SIR, SAR and SDR.

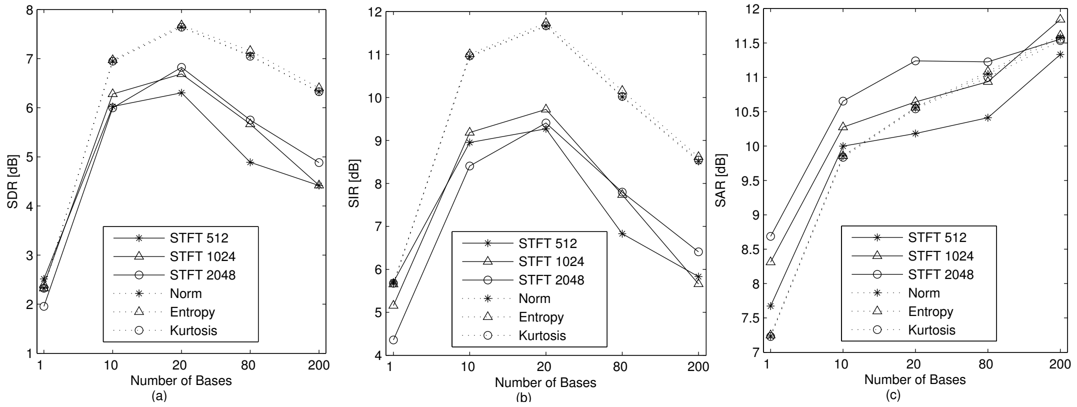

Fig. 2. The separation results in terms of (a)SDR, (b) SIR, (c) SAR for various number of bases per each source.

The separation is performed using the method described in Section 4 by adaptively choosing the time-frequency resolution. We calculate the STFT with three different window sizes: 512, 1024 and 2048 samples. Time-frequency magnitudes are calculated on the same grid by zero padding windowed signals and using equal STFT analysis hops of length 256 samples for every resolution. We performed NMF separation using the number of bases as $R = [1, 10, 20, 80, 200]$. We averaged the separation results from a set of five runs of five randomly selected pairs of male/female speakers from the TIMIT database. To obtain the optimal resolution at every point (f, t) of the time-frequency plane we consider a rectangular area Ω around this point with various sizes. The length of the rectangular grid is selected as the combinations of $Q = \{3, 53, 103\}$ time frames and $P = \{3, 53, 103\}$ frequency bins around the point of interest. We couldn't see significant difference if we change the size of the rectangular area Ω, but we get the slightly best results when we use 3 time frames and 103 frequency bins around each time-frequency bin for computing the sparsity. In Fig. 2, the SDR, SIR and SAR results obtained by the proposed method are displayed, respectively. The horizontal axis displays the number of bases per each source. In the figure, the results obtained by fixed time-frequency resolutions are also plotted to show the improvement by mixing the fixed time-frequency resolution results in an adaptive way. As it can be seen from the figure, the proposed method increases the separation quality by 1-3 dB in terms of SDR and SIR for $R \geq 10$. On the other hand, the SAR results seem to be the average of the SAR values obtained by the fixed time-frequency resolutions for $R \geq 20$. The results obtained by using different sparsity measures are displayed for comparison. We observe that if we compute the sparsity using entropy measure, we get a slightly better separation performance.

6. CONCLUSION

In this paper, we have presented an adaptive time-frequency resolution supervised method for separating known types of sounds from a single observation. We observed an improvement of 1-3 dB in terms of SDR and SIR relative to the fixed time-frequency resolution separation results. The future work will address extension of the method into convolutive methods and decreasing the computational complexity.

7. REFERENCES

[1] 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.

[2] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, January 2007.

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

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inform. Syst.*, vol. 13, pp. 556–562, 2000.

[5] N. Hurley and S. Richard, "Comparing measures of sparsity," in *IEEE Workshop on Machine Learning for Signal Processing*, 2008, pp. 55–60.

[6] T. Painter and A. Spanias, "A review of algorithms for perceptual coding of digital audio signals," in *13th International Conference on Digital Signal Processing*, 2-4 July 1997, vol. 1, pp. 179–208.

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