International Journal of Communication Networks and Information Security
ISSN: 2073-607X, 2076-0930
Volume 15 Issue 02 Year 2023

Non-Negative Matrix Factorization Based Single Channel Source Separation

SANTOSH KUMAR S*
Research Scholar, School of ECE, REVA University, Bengaluru, Department of ECE, Sri Venkateshwara College of Engineering, Bengaluru, India
reachsun@gmail.com
BHARATHI S H
Professor, School of ECE, REVA University, Bengaluru, India
bharathish@reva.edu.in

Article History

Received: 01 March 2023
Revised: 18 April 2023
Accepted: 29 August 2023

Abstract

The significance of speech recognition systems is widespread, encompassing applications like speech translation, robotics, and security. However, these systems often encounter challenges arising from noise and source mixing during signal acquisition, leading to performance degradation. Addressing this, cutting-edge solutions must effectively incorporate temporal dependencies spanning longer periods than a single time frame. To tackle this issue, this study introduces a novel model employing non-negative matrix factorization (NMF) modelling. This technique harnesses the scattering transform, involving wavelet filters and pyramid scattering, to compute sources and mitigate undesired signals. Once signal estimation is achieved, a source separation algorithm is devised, employing an optimization process grounded in training and testing approaches. By quantifying performance metrics, a comparative analysis is conducted between existing methods and the proposed model. Results indicate the superior performance of the suggested approach, underscored by these metrics. This signifies that the NMF and scattering transform-based model adeptly addresses the challenge of effectively utilizing temporal dependencies spanning more than a single time frame, ultimately enhancing speech recognition system efficacy.

Keywords: Automatic Speech Recognition, Matrix Factorization, Neural Network, Source Mixing, Wavelet Transform.

1. Introduction

In the past few years, significant focus has been directed towards investigating single-channel speech separation (SCSS) in a variety of research works. Its potential uses encompass domains related to speech signal processing, including Automatic Speech Recognition (ASR), speech separation, speech enhancement, and speech recognition systems. SCSS presents a unique challenge due to its confinement to a single microphone, aiming to extract specific target signals. The task of speech separation is categorized into three groups based on the available prior information [1]: fully supervised, partially supervised, and unsupervised scenarios. In the fully supervised approach, both speech and noise serve as inputs to generate corresponding dictionaries during training, which are then utilized for speech data computation. These learned dictionaries are typically fixed during separation, where activation is devised to reconstruct the original signal, an essential step for more

accurate signal reconstruction. The semi-supervised method involves instances of missing speech and noise data, requiring the separation stage to create the missing dictionary. Finally, the unsupervised model confronts the challenge of separating sources with unidentified speech sounds, making the separation process considerably intricate.

The process of extracting specific audio signals or sources from a blend of multiple sounds is known as speech separation. This becomes particularly challenging when the number of sources in the mixture exceeds the observable number of channels. Numerous potential solutions have emerged from research on this topic [2]. One approach involves modelling sources based on their spatial distinctions [3]. Leveraging variations in spatial characteristics, such as location or arrival direction, enables the separation of sources from the mixture. Another avenue explores the redundancy inherent in sources [4]. By capitalizing on the likelihood of sources carrying similar or overlapping information, separation algorithms can distinguish and segregate these redundant components. Sparsity-based methods have also gained prominence for speech separation [5]. These methods exploit the tendency for speech signals to exhibit sparsity in certain transform domains, such as the time-frequency domain. Encouraging sparsity in separated sources proves effective in achieving successful separation.

Non-negative matrix factorization (NMF) [6] has emerged as a recent solution for addressing speech separation challenges. NMF stands as an unsupervised technique for learning dictionaries, forming the basis of numerous sound separation methods[7]. It finds utility across diverse applications [8], [9], being widely adopted in both separation and training phases. NMF's effectiveness extends to solving research issues in speech enhancement [10], [11], speech separation [12], instrument segmentation [13], and even image processing [14]. NMF operates by breaking down an input signal matrix into a combination of non-negative coefficients and elements. The enforced non-negativity constraint ensures the absence of negative factors in the factorization. In the realm of single-channel source separation, a dictionary matrix, often sparse, is trained using clean or noisy speech data. This dictionary matrix usually contains numerous zero elements. Various strategies have been introduced to enhance NMF's efficacy in speech separation by incorporating additional constraints. For instance, the sparse constraint could penalize non-sparse vectors, encouraging sparsity within the factorization. These constraints aim to bolster the quality of separated speech signals and achieve more precise source separation.

The utilization of NMF in speech separation underscores its potential as a robust tool for disentangling mixture signals into their elemental sources. Through harnessing the non-negativity constraint and integrating supplementary constraints, NMF-based approaches exhibit promising outcomes in enhancing speech separation performance. Approaches such as the mixture of local dictionary [10], graph regularization, or utilization of NMF with β-divergence [12] have been employed. The NMF-based speech separation model is depicted in Figure 1 below.

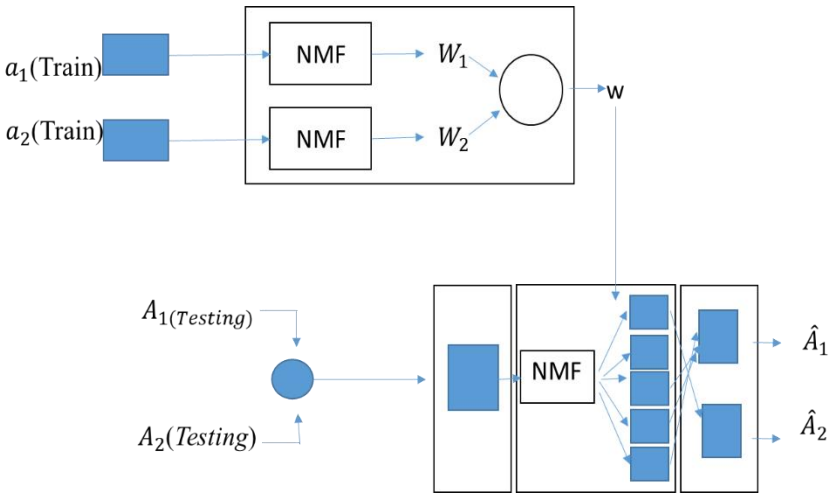

Fig 1. Architecture of NMF-based Speech Separation Model

The non-linear analysis operator and the synthesis op of the NMF-based speech separation technique can be thought of as a chain. Short Time Fourier Transformation (STFT) is used to convert the signal into a time-frequency representation, which is then used to create a feature space. Then, the dictionary is applied to obtain non-overlapping in the feature space. Finally, by inverting this non-linear representation of the separation is obtained, which is considered to be a crucial factor in

producing an efficient design/technique. As a general rule, the magnitude of STFT is invariant to changes in the phase; as a result, the NMF is freed from the burden of learning about the unsuitable variability. This occurs at the expense of inverting the estimates that do not overlap in the feature space, a process commonly referred to as the phase recovery problem [15] ,[16]. To address this, [17] [18] presented a technique for speech enhancement and musical source separation [19], and a deep learning technique has been adopted to compute magnitude source spectrograms [20] or to compute time-frequency masks [21].

Faced with challenges in solving the source separation problem, one needs to either develop its own model or choose an existing methodology to solve the problem at stake. This selection process depends on the kind of source and mixtures that need to be separated, which often leads to compromise between implementation complexity [22] and source separation quality. Once a methodology is selected, finding the objective of tuning parameters is challenging, which is driven by experience.

This model's goal is to demonstrate that employing a stable and resilient multi-resolution representation of the sample can be of assistance to the source separation methods when applied in discriminative contexts [23],[24],[25]. In this body of work, a deep representation is considered by the application of wavelet scattering pyramids, which generate information at various temporal resolutions [26]. In addition to this, it defines the parameter, which is becoming less expanding. The Constant Q Transforms (CQT) [27] could be thought of as having been simplified as a result of this. Scattering transformation, which can be applied to audio signals and has shown great performance in a variety of classification tasks, can be used to produce discriminative features that comprise longer temporal data [28]. These features can be created using scattering transformation. For the case of source separation, we must preserve temporal discriminability and, at the same time, capture long-range temporal representation [29]. The discriminative training model presented in this work shows better performance than state of the art model.

The following is an outline of the contribution that the research study has made:

- This work presented a discriminative training approach for solving speech separation problem.
- The model demonstrates that making use of a stable and resilient multi-resolution representation of the sample can be of assistance to the source separation methods in scenarios requiring discrimination.
- The model achieves significant performance over state of art models.

The organization of the remainder of this paper is outlined as follows. In Section II, we provide a comprehensive description of the novel algorithm devised for the purpose of source separation. To comprehensively evaluate the efficacy of our proposed model, we conduct an experimental study in the preceding section, where we compare its performance against that of the current state-of-the-art technique. Finally, in the concluding section, we thoroughly analyze and interpret the results obtained while also delineating potential avenues for future research exploration.

2. Related works

Here, we present an efficient audio source separation using a discriminative training approach.

2.1 Problem formulation and System Model

Optimization is said to be the key to solving the speech separation problem. Let the voice signals of the first and second users be represented by $a_1(n)$ and $a_2(n)$, where n is a discrete-time index. We also take into account that not all speech signals are created equal.

$$\sum_n a_1^2(n) = \sum_n a_2^2(n) = 1, \tag{1}$$

and that $a_1(n)$, $a_1(n)$ are zero mean and are statistically independent. Let's consider a set of speech signals a_1 and a_2 from two speakers, and temporal signal $b(n)$ observed from these two speech signals is obtained as follows

$$b(n) = a_1(n) + a_2(n) \tag{2}$$

and our objective is to compute $\hat{a}_x$. Here, we use an NMF-based source separation technique. Most of the traditional model adopts a non-negative time-frequency of $b(t)$, such as power (or magnitude), which is denoted as $\beta(b) \in \mathcal{K}^{i \times j}$, where i and j represent temporal frames and frequency, respectively.

This transformation is considered to be a non-linear analysis function such as the Short-Time Fourier Transform (STFT). Let $\mathcal{Z}(\ell, \omega)$ be the STFT of $a_1(n)$, where ℓ is the temporal frame index and $0 \leq \omega \leq \mathcal{W} - 1$ is the frequency index. The $\frac{\mathcal{W}}{2} + 1$ dimensional vector is denoted by z_ℓ, whose ω^{th}element is

$$z_\ell^{\omega} = \log|\mathcal{Z}(\ell, \omega)|; \qquad \omega = 0,1, \ldots, \mathcal{W}/2. \tag{3}$$

Therefore, z_ℓ and $\bar{z}_\ell$ are the log-spectrum of $a_1(n)$, and $a_2(n)$, respectively.

2.2 Non-Negative Matrix Factorization

NMF-based source separation techniques have attained wide interest in recent times. The NMF-based approach finds the non-negative functions $T_x \in \mathcal{K}^{y \times j}$, x representing the best matching component in dictionaries [15] $\mathbb{L}_x \in \mathcal{K}^{i \times y}$. This is solved using the following solution as

$$\min_{\bar{a}_x, T_x \geq 0} \sum_{x=1,2} \mathbb{L}(\beta(\bar{a}_x)|\mathbb{L}_x T_x) + \mu\mathcal{K}(T_x) \qquad such\ that\ b = \bar{a}_1 + \bar{a}_2. \tag{4}$$

The term in Eq. (4) computes the variation among the input data, and the estimated channel and choices of $\mathbb{L}$ are the squared Euclidean distance, the Itakura-Saito divergence, and the Kullback-Leibler divergence. The second term in Eq. (4) improves the preferred structure of functions. This process is processed using the regularization function $\mathcal{K}$, whose relative parameter is controlled by parameter μ. Here, we consider $\mathbb{L}$ reweighted squared Euclidean distance and ℓ_1 standard as regularization function $\mathcal{K}$.

The problem in Eq. (4), can be further minimized by alternating gradient descent among $\bar{a}_x$ and t_x. However, the phase recovery problem can be solved by fixing t_x and minimizing it with respect to $\bar{a}_x$, which requires locally inverting the transform β. Therefore, the first separation from feature space is obtained by applying classic NMF as follows

$$\min_{T_x \geq 0} \mathbb{L}\left(\beta(a)\middle|\ \Sigma_{x=1,2}\, \mathbb{L}_x T_x\right) + \mu \sum_{x=1,2} \mathcal{K}(T_x) \tag{5}$$

Once solving activation functions are completed, the spectral envelope is computed as $\beta(\hat{a}_x) = \mathbb{L}_x T_x$, and the phase recovery problem can be solved by applying soft masks to filter the m0ixture signal. That is, $\beta(b) = |M\{b\}| \in \mathbb{Q}^{i \times j}$ is a complex matrix with respect to STFT. The computed pure signals are acquired by filtering the mixture,

$$\hat{a}_x = \frac{1}{S}\{I_x \cdot M\{b\}\}, \qquad \text{with} \quad \frac{\beta(\bar{a}_x)^h}{\Sigma_{g=1,2}\beta(\bar{a}_x)^h}, \tag{6}$$

where h represents the smoothness of the mask. The objective is to build a dictionary such that Eq. (5) minimizes the reconstruction considering the ground truth separation,

$$\min_{\mathbb{L}_1 \geq 0, \mathbb{L}_1 \geq 0} \mathbb{L}(\beta(a_1)|\mathbb{L}_1 T_1^s) + \varphi\mathbb{L}(\beta(a_2)|\mathbb{L}_2 T_x^s) \tag{7}$$

where φ is a controlling relative factor of source recovery and T_x^s are the solution of Eq. (5).

2.3 Wavelet filter transformation

In our work, we adopt a complex wavelet with a quadratic phase whose Fourier transform fulfils $\mathbb{F}\alpha(\gamma) \approx 0$ for $\gamma < 0$. We consider that bandwidth is of the order of $\mathbb{Q}^{-1}$ and the center of frequency of $\mathbb{F}\alpha$ is one, which is computed by increasing

$$\alpha: \alpha_\lambda(n) = \lambda\alpha(\lambda n) \tag{8}$$

such that when $\lambda = 2^{y/\mathbb{Q}}$ and therefore, $\mathbb{F}\alpha_\lambda(\gamma) = \hat{\alpha}(\lambda^{-1}\gamma)$. The set of indices denoted by $\lambda = 2^{y/\mathbb{Q}}$ over the frequency support of the signal, where $y \leq Y_1$, is elegantly represented as by τ. This ultimately leads to the establishment of a filter bank that exhibits a consistent parameter $\beta_1(n)$ for the bands within each octave (Y_1). For the purpose of analysis, let's define $\beta_1(n)$as the bandwidth of a low-pass filter with a value of 2^{-Y_1}.

$$J_a = \{a \times \alpha_\lambda(n), a \times \alpha_\lambda(n)\}_{\lambda \in \tau}. \tag{9}$$

Given that the bandwidth of each filter is constrained to a maximum of $\mathbb{Q}^1$, it becomes feasible to implement a down-sampling process on its resulting output, employing a stride of $\mathbb{Q}$. This approach enables us to efficiently reduce the computational load while retaining essential information in the analysis. Consequently, by strategically down-sampling the outcomes, we strike a balance between computational efficiency and preserving key features within the signal processing procedure.

To retain temporal locality within the analysis of wavelet coefficients, a crucial sampling rate is

embraced instead of a uniform bandwidth smoothing kernel applied across all layers. This pivotal sampling rate ensures adequate temporal information capture during signal sampling. To achieve this, the complex phase of J_a wavelet coefficients is initially removed through a complex modulus nonlinearity. This step simplifies coefficient representation while safeguarding magnitude data. Subsequently, the initial layer coefficients are organized as nodes within the first-tier tree level. This organization streamlines subsequent down-sampling. Each tree node signifies a coefficient at a distinct scale and temporal-frequency position. Down-sampling occurs at the critical sampling rate of this layer, designated as layer δ_1. This rate corresponds to the reciprocal of the highest bandwidth within the filter bank employed for wavelet transformation. This approach retains vital temporal details while mitigating data size by employing the critical down-sampling rate.

$$\left|J^1\right|a = \left\{a_x^1\right\}_{x=1\ldots1+|\tau|} = \{a \times \beta_1(\delta_1 j), |a \times \alpha_\lambda(\delta_1 j)|\}_{\lambda\in\tau}. \tag{10}$$

The coefficients within the initial layer play a crucial role in offering localized feedback both in terms of frequency and temporal aspects. However, this advantage comes at a certain cost, manifested as a trade-off concerning the parameter $\mathbb{Q}$. Following the down-sampling step, each of the resultant signals undergoes a transformation into a novel wavelet filter bank. These transformed signals are then subjected to the operation of taking the complex modulus of their oscillatory components. This adjustment aims to enhance the accuracy of the representation by mitigating potential noise or fluctuations. The primary objective of this operation is to bolster the robustness of the overall representation. As such, we undertake the process of simplifying the complex filter bank by condensing it into a pair of conjugate mirror filters, namely $\{\beta_2, \alpha_2\}$. These filters are strategically designed to respectively capture the low and high-frequency components of the discrete signal emanating from above the hierarchical structure. This simplification is introduced with the intention of enhancing computational convenience while still retaining vital signal characteristics that are essential for further analysis,

$$\left|J^2\right|a = \left\{a_x^1 \times \beta_2(2j), \left|a_x^1 \times \beta_2(2j)\right|\right\}_{x=1\ldots\left|J^1\right|}. \tag{11}$$

As a consequence, each successive layer within this framework contributes novel feature vectors, albeit at a reduced temporal resolution. This is achieved by meticulously considering only the coefficients that have been subjected to the nonlinearity operation for layers $i \le i_\uparrow$. Given that the energy of the signal experiences rapid decay, it becomes plausible to iteratively apply the transformationω times, as needed, until a specific temporal condition is met—namely, $N = 2^\omega \delta_1$. In the event that the wavelet filters are thoughtfully selected to embody a non-expansive vector, a notable consequence ensues: each layer inherently encapsulates a metric that progressively diminishes. This intriguing property is a result of the inherent structure of the chosen wavelet filters, which imparts a cumulative reduction in the spatial metric as we traverse through the layers of the analysis,

$$\left\|\left|J^\omega\right|a - \left|J^\omega\right|\bar{a}\right\| \le \left\|\left|J^{\omega-1}\right|a - \left|J^{\omega-1}\right|\bar{a}\right\| \le \|a - \bar{a}\|. \tag{12}$$

Therefore, each layer gives new feature vectors at a lower temporal resolution. Then, lastly, we acquire a tree of different representations, $\beta_y(a) = \left|J^y\right|a$ with $y = 1, \ldots, \omega$.

2.4 Source Separation Model

Here, we show how the pyramid scattering feature can solve the source separation problem of Eq. (2). We use a discriminative training approach. First, we obtain models of each user using wavelet scattering pyramid features. Transformation of each layer produces feedback with different discriminating trade-offs.

Let us consider two distinct sources labelled A_1 and A_2. To simplify the discussion, we will focus on a feature denoted as $\beta^y(a_x)$,, where $y = 1,2$, $x = 1,2$ with $a_x \in A_x$, belonging to the set A_x. These features are obtained by extracting the localized scattering characteristics from two separate resolutions at their respective sampling rates. This process ensures that the features are tailored to capture the unique qualities of the signals originating from A_1 and A_2 while considering their distinct temporal resolutions. In light of this, it becomes evident that the feature β_1 provides a more pronounced level of discriminative and localized feedback compared to β_2. In our methodology, we proceed by training independent models for each individual audio sample. When working with a training Set A_x^n extracted from each audio sample, we delve into a process involving Non-Negative Matrix Factorization (NMF) applied to the features $\beta_y(a_x^n)$,.

This approach enables us to uncover the inherent patterns and components present within each set of features, effectively capturing the underlying structure of the audio samples in a manner conducive to subsequent analysis and interpretation. By training these distinct models and applying NMF to the localized scattering features, we attain a comprehensive understanding of the characteristics intrinsic to each audio sample. This forms a foundational step for further exploration, manipulation, and insight generation,

$$\min_{\mathbb{L}_x^y, T_x^y \geq 0} \sum_{a_x \in A_x} \frac{1}{2} \left\| \beta^y(a_x) - \mathbb{L}_x^y T_x^y \right\|^2 + \lambda_x^y \left\| T_x^y \right\|_1 \tag{13}$$

where λ_x^y represents the sparsity reconstruction trade-off controlling factor in space coding. For a given time $b = a_1 + a_2$. We compute $\bar{a}_1$, $\bar{a}_2$ as the solution of

$$\min_{\bar{a}_1 + \bar{a}_2 = y,\, T_x^y \geq 0} \sum_{x=1,2} \frac{1}{2} \left\| \beta^1(\hat{a}_x) - \mathbb{L}_x^1 T_x^1 \right\|_2^2 + \lambda_x^1 \| T_x \|_1 + \frac{1}{2} \left\| \beta^2(\bar{a}_x) - \mathbb{L}_x^2 T_x^2 \right\|_2^2 + \lambda_x^2 \left\| T_x^2 \right\|_1 . \tag{14}$$

The Eq. (14) under linear constraint may suffer due to coupled phase recovery problems. To address we use the greedy approach, which approximates the unidentified complex phase using the phase of J_{1b} and $J_2 \left| J_{2y} \right|$ respectively. We solve by using a stronger version of linear constraint $b = a_1 + a_2$, namely

$$\left| J^1 b \right|^2 = \left| J^1 a_1 \right|^2 + \left| J^1 a_2 \right|^2, \tag{15}$$

Therefore, destructive interference is trivial. In next section, the experiment analysis of the proposed speech separation method is evaluated.

3. Methodology and Results

The experiments were conducted on a system running the Windows 10 operating system, equipped with an Intel I-5 class processor, 16GB RAM, and a dedicated NVIDIA CUDA GPU with 4GB RAM. The proposed NMF-based Speech Separation model was implemented using MATLAB. The experimental study employs the TIMIT dataset [23] [24], which follows a test-train division. In the case of multi-speaker scenarios, the model was trained using mixed speech samples of both male and female speakers. Testing was then performed using samples not included in the training dataset. Signal mixing and resampling were conducted according to the parameters specified in Table 1. The dataset comprised samples from three male and three female users. For training, 500 clips were employed, while testing was executed on 12 distinct users, with an equal split of 6 male and 6 female users, encompassing various combinations of samples.

Table 1. Simulation Parameter

Parameter	Value
Sampling-Frequency	16 Khz
Signal Mixing	0 db
Resolution Adjustment	32
Sampling Time	2048
Normalization Constant	1e-3
Number of Iteration	70

The performance of the proposed model is evaluated by comparing it to state-of-the-art models referenced as [22], [23], [24]. This evaluation involves the use of metrics such as Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), and Source to Artifact Ratio (SAR).
These metrics provide quantitative measures of the quality of the separated speech signals. The computation of these metrics is typically performed as follows:

1. Source to Distortion Ratio (SDR): SDR measures the ratio of the power of the source signal to the power of the distortion or interference introduced during the separation process. It can be computed by comparing the estimated source signal to the reference or ground truth source signal. The SDR is calculated as the average of the logarithm of the ratio of the estimated source signal's energy to the energy of the interference or distortion:

$$\mathrm{SIR} = 10\log\frac{\sum_t y_{is}^2(t)}{\sum_{i\neq j}\sum_t\sum_t y_{is}^2(t)} \tag{16}$$

2. Source to Interference Ratio (SIR): SIR quantifies the ratio of the power of the source signal to the power of the interference, including residual noise and artifacts. It provides an indication of how well the separation model can suppress unwanted components. Similar to SDR, SIR is computed as the average logarithmic ratio of the estimated source signal's energy to the energy of the interference:

$$\mathrm{SDR} = 10\log\frac{\sum_t s_{is}^2(t)}{\sum_t\left(s_k(t) - \mathbb{C}y_{is}(t-\beta)\right)^2} \tag{17}$$

y_{is}^2 is the estimated signal of $s(t)$, $\mathbb{C}$ is constant to compensate for the difference in amplitudes, and the phase difference between input and output is given by β.

3. Source to Artifact Ratio (SAR): SAR measures the ratio of the power of the source signal to the power of any artifacts introduced during the separation process. It represents the ability of the model to remove artifacts or distortions. SAR is computed in a similar way to SDR and SIR by comparing the energy of the estimated source signal to the energy of the artifacts:

$$SAR = 10\log_{10}\frac{\left||target + interference + noise|\right|^2}{\left||artifact|\right|^2} \tag{18}$$

In the approach that has been described, a conventional NMF model is utilized. This model has 1024 samples, and 50% of them overlap. Four hundred atoms are utilized in the multi-speaker model for the purpose of dictionary selection. These parameters were determined through the process of cross-validation on a selection of clips that were kept apart from the training set to serve as the validation set. The performance of our model over a state of the art model is evaluated in terms of SDR, SIR, and SAR which is cumulated in Table 2.

Table 2. Multi-Speaker Model Performance Comparison

	SDR (dB)	**SAR**	**SIR**
NMF [24]	6.008	7.624	8.722
DNN [25]	7.7	8.07	11.53
$DNNcIRM$ [24]	7.4	7.5	12
$DNNcIRM - dis$ [24]	7.67	7.89	12.83
$KL + TC + Weiner$ [26]	N/A	N/A	7.1152
MPS [27]	8.25	9.67	11.45
$MPS - WTA$ [27]	8.17	9.61	11.33
Proposed NMF	**9.491**	**10.396**	**13.594**

The result obtained in Table 2 shows that the proposed NMF-based speech separation model outperforms a state of the art technique [22],[23],[24],[25]. The Proposed NMF achieves around 1.82 dB, 2.506 dB and 0.744 dB gain compared to $DNNcIRM - dis$ [23] in terms of SDR, SAR and SIR, respectively. For further case studies, the performance of the proposed NMF-based speech separation model is compared with $KL + TC + Weiner$ [25], and outcomes show a SIR gain of 6.478 dB is achieved. The Proposed NMF achieves around 1.321 dB, 0.786 dB and 2.264 dB gain compared to $MPS - WTA$ [22] in terms of SDR, SAR and SIR, respectively. The overall result obtained in TABLE II shows that the proposed NMF -based speech separation model outperforms all state of art techniques [22], [23], [24], [25] in terms of SIR, SAR and SDR.

4. Conclusion

In this research, the authors introduce a novel approach to source separation through the utilization of non-negative matrix factorization (NMF). Practical implementation of this method involves signal processing techniques, such as resampling input data and merging two audio samples. Subsequently, a wavelet filter bank and pyramid scattering come into play to disentangle the components of the mixed signal. Finally, a source separation algorithm is employed to effectively untangle the sources, eliminating their overlap. The performance of the proposed NMF model was empirically assessed and juxtaposed with prevailing best practices. Evaluation metrics like Source to Artifact Ratio, Source to Distortion Ratio, and Source to Interference Ratio were employed. The results affirm that the suggested NMF model outperforms state-of-the-art methods, exhibiting superior Signal-to-Noise Ratio (SNR), Source to Artifact Ratio (SAR), and Source to Interference Ratio (SIR) by average margins of 1.958 dB, 2.002 dB, and 2.884 dB, respectively. These improvements underscore the elevated effectiveness of the proposed model compared to prior methodologies.

5. Acknowledgments

Authors acknowledge the support from REVA University for the facilities provided to carry out the research.

References

[1] Y.V. Koteswararao and C.B. Rama Rao, “Single channel source separation using time–frequency non-negative matrix factorization and sigmoid base normalization deep neural networks”, *Multidimensional Systems and Signal Processing*, vol. 33, no. 3, pp.1023-1043, 2022.

[2] N.C. Nag, and M.S. Shahm “Investigating Single Channel Source Separation Using Non-Negative Matrix Factorization and Its Variants for Overlapping Speech Signal. In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-6). IEEE, January 2019.

[3] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M.A. Mahjoub and C. Cleder, Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication*, vol. 114, pp.22-35, 2019.

[4] Z. Chen, J. Droppo, J. Li and W. Xiong, “Progressive joint modeling in unsupervised single-channel overlapped speech recognition”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp.184-196, 2017.

[5] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J. P. Xu, and G. Z. Tan, “Speech emotion recognition based on feature selection and extreme learning machine decision tree,” *Neurocomputing*, vol. 273, pp. 271–280, Jan, 2018.

[6] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, “From blind to guided audio source separation: How models and side information can improve the separation of sound,” IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 107–115, 2020.

[7] A. Jourjine, S. Rickard, and O. Yılmaz, “Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,” in Proc. Of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), vol. 5, pp. 2985–2988, 2000.

[8] P. D. O’Grady, B. A. Pearlmutter, and S. T. Rickard, “Survey of sparse and non-sparse methods in source separation,” International Journal of Imaging Systems and Technology, vol. 15, no. 1, pp. 18–33, 2005.

[9] D.F. Rosenthal, H.G. Okuno, H. Okuno and D. “Rosenthal, Computational Auditory Scene Analysis: Proceedings of the Ijcai-95 Workshop (1st ed.)”, *CRC Press*, 1998.

[10] J.-L. Durrieu, G. Richard, B. David, and C. F'evotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 3, pp. 564–575, 2010.

[11] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” Advances in Neural Information Processing Systems, Cambridge, MA, USA: MIT Press, vol. 13, 2000.

[12] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, “From blind to guided audio source separation: How models and side information can improve the separation of sound,” IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 107–115, 2014.

[13]T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," Signal Processing Magazine, IEEE, vol. 32, no. 2, pp. 125–144, 2015.

[14]A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," IEEE Transactions on Signal Processing, vol. 62, no. 16, pp. 4298–4310, 2014.

[15]P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-Euclidean sparse NMF via bilevel optimization with applications to speech enhancement," Joint Workshop on Hands-free Speech Communication and Microphone Arrays, pp. 11 - 15, May 2014.

[16]A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP), 2011.

[17]J. T. Chien and H. L. Hsieh, "Bayesian group sparse learning for music source separation," EURASIP Journal on Audio, Speech, and Music Processing, no. 18, pp. 1–15, 2013.

[18]J. L. Roux, F. Weninger, and J. R. Hershey, "Sparse NMF – half-baked or well done?," *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, *11*, pp.13-15.Mar 2015.

[19] R.W. Gerchberg, "A practical algorithm for the determination of plane from image and diffraction pictures", *Optik*, vol. 35, no. 2, pp.237-246, 1972.

[20]C. F'evotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3158-3162). IEEE, 2013.

[21]J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. W¨ollmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM + TUT + KUL approach to the 2nd CHiME challenge: Multi-stream ASR exploiting BLSTM networks and sparse NMF," Proc. of CHiME-2013, pp. 25–30, 2013.

[22]F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, pp. 577–581, 2014.

[23]E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013, pp. 162–167.

[24]J. L. Durrieu, G. Richard, B. David, and C. F'evotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 3, pp. 564–575, 2010.

[25]E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 3734-738.

[26]D. S. Williamson, Y. Wang, D. Wang, "Complex Ratio Masking for Monaural Speech Separation. IEEE/ACM Trans Audio Speech Lang Process", vol. 24, no. 3, pp. 483-492, 2016

[27]Z. Miao, X. Ma, and S. Ding, "Phase constraint and deep neural network for speech separation", *In Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II 14* (pp. 266-273). Springer International Publishing, 2017.

[28]S. Abdali, B. NaserSharif, "Non-negative matrix factorization for speech/music separation usingsource dependent decomposition rank, temporal continuity term and filtering", *Biomedical Signal Processing and Control*, Vol. 36, pp. 168-175, 2017.

[29]M. Kim, P. Smaragdis, and G.J. Mysore, Efficient manifold preserving audio source separation using locality sensitive hashing. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 479-483). IEEE, April 2015.

[30]K. Minje and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," IEEE Signal Processing Letters, vol. 22, no. 3, pp. 293-297, March 2014.