IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 10, 2020, accepted July 13, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010342

Hybrid Source Prior Based Independent Vector Analysis for Blind Separation of Speech Signals

JUNAID BAHADAR KHAN[1], TARIQULLAH JAN[1], RUHUL AMIN KHALIL[1], (Graduate Student Member, IEEE), AND ALI ALTALBE[2]

[1]Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Engineering and Technology Peshawar, Peshawar 25120, Pakistan

[2]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Junaid Bahadar Khan (jbkh08@gmail.com)

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors therefore, acknowledge with thanks DSR for technical and financial support.

ABSTRACT Blind Source Separation (BSS) application is a delinquent issue in a complex reverberant environment with changing room geometric dimensions and an increasing number of speech sources. The BSS application issue is determined by the independent component analysis that usually manipulates higher-order statistical approaches. However, the permutation between desired speech sources remains a challenging issue for BSS applications. The permutation problem is been rectified by Independent Vector Analysis (IVA) for BSS applications in the frequency domain. The performance dependency of the IVA approach solely relies on the selection of appropriate source-prior to preserve the inter-frequency dependencies between the same speech source amongst different frequency bins. Therefore, a hybrid model for the IVA method is presented, which comprises of multivariate generalized Gaussian and super-Gaussian distribution source priors to model low as well as high amplitudes speech signals. The weights of the hybrid model between multivariate Gaussian and generalized Gaussian are assigned in accordance to the energy of the observed non-stationary speech mixture signal. In the simulations, different speech mixtures are generated from various speech sources by simulated room model. The proposed approach evaluates the blind separation performance in terms of signal-to-distortion ratio (SDR) and is compared with well-known BSS methods. The results show an improvement of the proposed methodology for non-stationary speech signals over the state-of-the-art IVA models having a fixed source prior.

INDEX TERMS Blind source separation (BSS), convolutive speech mixture, independent vector analysis.

I. INTRODUCTION

Human listeners show the ability to separate the desired speech signal from complex auditory speech mixture, i.e. cocktail party environment [1]. However, humans with hearing loss suffer significant intelligibility of desired speech signal in a noisy reverberant environment. Amplifying the receiving speech mixture cannot solve the intelligibility of the desired speech signal as it amplifies the targeted as well as the interfering speech signals. Therefore, separation of the target speech signal from a complex speech mixture is a challenge for speech processing machines [2]. The separation, and to preserve the intelligibility of the desired speech become more challenging in a noisy reverberant environment [3], [4]. The cocktail party problem is reviewed for years and distinctive solutions are provided by the researchers to mitigate the problem [5]–[7]. Despite the current solution provided, still cocktail party problem remain a scientific challenge for the researchers and demand further scientific research efforts. The solutions to this problem will have a vast impact on speech processing applications such as digital hearing aids, audio speech sensors, automatic speech recognition (ASR), hands-free communication devices, binaural telephone headsets, and acoustic surveillance systems [8]–[11].

Blind Source Separation (BSS) framework is the state-of-the-art method for the separation of target speech from a

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

mixture of speech signals. In BSS applications, the desired targeted speech signal is separated from the speech mixture without prior knowledge of the speech signal and the mixing process. The popular tool used in BSS is independent component analysis (ICA) [12]–[14]. ICA approach uses higher-order statistical models for separating the sources from the mixed speech signals with the assumption of statistically independent speech source signals. In real-time environment, the performance of ICA model is enhanced for convolutive reverberant speech mixture by combining multistage approaches with ICA. It is combined with Binary mask to increase the separation gain [15]. Furthermore, cepstral smoothing is combined with ICA and Binary masking to reduce the musical noise caused by time-frequency masking [16]. The ICA method is also used to combine antenna array to sense different sources remotely [17].

In the time domain, the convolutive reverberant mixture becomes computationally complex and time-consuming which results in performance degradation of speech processing systems. The computational cost of ICA algorithm is reduced by online recursive ICA model [18]. Furthermore, this problem is overcome by transforming time-domain (TD) into frequency domain (FD) using short-time Fourier transform (STFT) approach [16], [19], [20]. However, the main problem of permutation exists across different frequency blocks [8]. Different proposals such as consistency of filter coefficients, consistency of spectrum and variational Bayesian are provided to resolve the issue [21], [22]. However, the solutions provided require pre and post processing, increasing the computational processing of ICA approach. Therefore, permutation problem is a challenging issue. This problem is rectified in the frequency domain by an approach known as Independent Vector Analysis (IVA). It avoids the permutation problem in the learning process without pre or post-processing [8]. IVA approach preserves the inter-frequency dependency of the same source among different frequency bins. The frequency components are assumed to be independent from different speech sources across the frequency bins [8].

The separation performance of IVA algorithm strongly depends on the selection of source prior, which accommodates the inter-frequency dependency amongst different frequency blocks of a source speech signals. The fundamental IVA algorithm uses multivariate Laplace distributions (MLD) as a source prior to couple higher-order dependencies [8]. In [23], the multivariate Gaussian distribution (MGD) source prior is utilized for recovering the desired speech signal from the mixture. However, both [8], [23] exploits only second-order statistics, and are unable to resolve high-order statistics.

The Performance of IVA is a further improved by using IVA algorithm with conjunction of ideal binary masking (IBM) and post processing by cepstral domain [24], [25]. In [26], a new family of multivariate distribution known as Kotz distribution is introduced, which is more flexible distribution to be used for source prior. This source prior has the capability of exploiting second order as well as higher order statistics. The high amplitudes in a voice mixture be better modeled by Student's T source prior for the IVA [27]–[29]. The heavy tailed nature of Student's T distribution enhances the separation performance significantly [30]. Mixed source prior such as, MGD and multivariate Student's T distribution are introduced to model the non-stationary nature of the observed mixture signal. The source priors switch between MGD and Student's T in accordance to the energy in each frequency bins of the observed speech mixture [31].

The state-of-the-art IVA method use only fixed source prior distribution models in the separation process of the BSS applications. This cannot better model the non-stationary nature of the observed speech mixture, resulting in the degradation of separation performance of the BSS application in real-time environment. Therefore, in this research work, a hybrid energy-based source prior is proposed, which enhances the robustness of the speech processing applications by adopting the IVA algorithm in accordance to the non-stationary nature of observed speech mixture. The proposed hybrid model comprises multivariate generalized Gaussian and super-Gaussian distributions for the IVA algorithm. The generalized Gaussian distribution with heavier tails will better model the higher amplitude of the source signals in the speech mixture, while the multivariate super-Gaussian distribution will extract other important information. The weights of the distributions in hybrid source prior are adapted based on the energy in each frequency bin of the observed mixture speech signal. More weight is given to generalize Gaussian source prior distribution if the frequency block of the observed speech mixture has high energy and vice versa. The proposed IVA model is simulated on stimulated Room Impulse Response (RIR). The simulation results show improvement in the IVA algorithm by the proposed methodology.

The organization of the paper comprises the following sections. Section II describes the independent vector analysis. Section III explains the multivariate source prior. Section IV describes the proposed hybrid source model. Results and discussion are provided in detail in Section V. The proposed model for source separation is investigated and its performance is evaluated in Section VI. Finally, conclusions are provided in Section VI, followed by future work.

II. INDEPENDENT VECTOR ANALYSIS

A cocktail party environment having M microphones and N number of speech sources. The observed mixture speech at each microphone can be mathematically expressed by the linear convolutive model as,

$$x_i(t) = \sum_{j=1}^{N} \sum_{\tau=0}^{T-1} h_{ij}(\tau) s_j(t-\tau); \quad i = 1, 2, \ldots, M \qquad (1)$$

in (1), the term $h_{ij}(\tau)$ depicts the impulse response of the room, that varies from j-th source toward i-th mixed signal. It should be noted that the mixed-signal i-th is having a

length of T, where the term $s_j(t)$ is the j-th source speech signal at time t. The computation outlay is further reduced by short-time Fourier transform (STFT) in converting the time-domain convolutional signal into multiplication in the frequency domain (FD). The time-domain speech signal is represented in FD compact form as,

$$X(K) = H(K)S(K) \tag{2}$$
$$\hat{S}(K) = W(K)X(K) \tag{3}$$

where, $X(K) = [x_1(k), x_2(k), \ldots, x_M(k)]^T$ is the observed mixture of speech signals, and $\hat{S}(K) = \hat{s}_1(k), \hat{s}_2(k), \ldots, \hat{s}_N(k)]^T$ is the estimated vector for speech signal. The vector for speech signal is expressed in frequency domain, where $(.)^T$ represents vector transpose. The terms $W(K)$ and $H(K)$ are the respective un-mixing and mixing matrices. Further, the index k depict the k-th frequency bin of FD-BSS model. In the proposed research, the number of receiving microphones and speech sources are considered equal i.e. $M = N = 2$.

The separation of multivariate speech sources from the observed multivariate mixture signal, a cost function will be defined for multivariate variables. Therefore, Kullback-Leibler (KL) divergence is used for the measurement of relative dependencies between the two functions, one having exact joint probability density function and the other having product of individual probability density function. It can be expressed as,

$$\begin{aligned} C &= KL(P(\hat{s}_1, \ldots, \hat{s}_N)||\prod_{i=1}^{N} q(\hat{s}_i)) \\ &= \text{const} - \sum_{k=1}^{K} \log|\det(W(k))| - \sum_{i=1}^{N} E \log q(\hat{s}_i) \end{aligned} \tag{4}$$

The dependencies between various speech sources are removed by minimizing the cost function preserved by each source vector. Therefore, for minimization of such dependencies, the gradient descent method is utilized to the cost function with respect to the un-mixing matrix $w_{ij}(k)$ [32].

$$\begin{aligned} \Delta w_{ij}(k) &= -\frac{\partial C}{\partial w_{ij}(k)} \\ &= \sum_{l=1}^{N} (I_{il} - E\varphi^k(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})\hat{s}_l^{(k)})w_{lj}^{(k)} \end{aligned} \tag{5}$$

where I and $\varphi^{(k)}(.)$ are the respective identity matrix and non-linear score function. The score function is expressed by,

$$\varphi^k\left(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)}\right) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})}{\partial \hat{s}_i^k} \tag{6}$$

Generally, online updates or batch update rules are used to update coefficients of the un-mixing matrix. The batch update rule is mathematically expressed as,

$$w_{ij}^{new}(k) = w_{ij}^{old}(k) + \eta \Delta w_{ij}(k) \tag{7}$$

where η is known as the rate of learning. By neglecting the expectation operation from (5) and updated at every sample time to achieve online update.

III. MULTIVARIATE SOURCE PRIORS

The speech signal dependencies in different frequency bins can be modeled by the probability density function (pdf). The IVA approach uses a super-Gaussian source prior. Mathematically, it is expressed as [8],

$$q(s_i) \propto \exp\left(-\sqrt{\sum_{k=1}^{K}\left|\frac{\hat{s}_i(k)}{\sigma_i(k)}\right|^2}\right) \tag{8}$$

in (8), $\sigma_i(k)$ represents variance of the i-th speech source at k-th frequency bin. From (6) the score function (non-linear) for the source is expressed as [8],

$$\varphi^{(k)}\left(\hat{s}_i(1), \ldots, \hat{s}_i(K)\right) = \frac{\hat{s}_i(k)}{\sqrt{\sum_{k=1}^{K}|\hat{s}_i(k)|^2}} \tag{9}$$

The IVA algorithm separation performance strongly relies on the score function, which is deduced from the source prior. The score function accommodates inter-frequency dependency. IVA approach performance is enhanced by selecting the appropriate source prior. In [8], each source of the covariance matrix is assumed as the identity matrix. Therefore, the second-order correlation is mitigated in different frequency bins. However, the multivariate Gaussian source prior introduced second-order correlation [23]. Though higher-order correlation is still missing amongst various frequency bins.

Furthermore, a multivariate generalized Gaussian distribution source prior is utilized to exploit a higher correlation in the frequency bins. Its heavier tails can model higher amplitudes and more robust to outliers. Moreover, the performance is further enhanced by introducing energy correlation for the source vector [33]. Therefore, extracting more dependent information between the frequency bins helps to improve the separation process. The multivariate generalize Gaussian distribution can be written as [33],

$$q(s_i) \propto \exp\left(-\left(\frac{(s_i - \mu_i)^\dagger \sum_i^{-1}(s_i - \mu_i)}{\alpha}\right)^\beta\right) \tag{10}$$

in (10), $(.)^\dagger$ is the Hermitian transpose, where the terms μ_i and $\sum_i$ are the respective mean and covariance of i-th source. β represents the shape parameter and α denotes the scaling factor. Assuming $\alpha = 1$, $\mu_i = 0$, then (10) becomes,

$$q(s_i) \propto \exp\left(-\left(\sum_{k=1}^{K}|s_i(k)|^2\right)^\beta\right) \tag{11}$$

using (6) the non-linear score function of multivariate generalize Gaussian source prior will become

$$\varphi^{(k)}(\hat{s}_i(1), \ldots, \hat{s}_i(K)) = \frac{2\beta\hat{s}_i(k)}{((\sum_{k=1}^{K}|s_i(k)|^2)^2)^{\frac{1-\beta}{2}}} \tag{12}$$

β is obtained by satisfying the following condition [33],

$$\frac{1-\beta}{2} = \frac{1}{2I+1} \tag{13}$$

$$\beta = \frac{2I-1}{2I+1} \tag{14}$$

where I represent positive integer and equal to 1. β should be considered less than $\frac{1}{2}$ to make the generalize Gaussian source robust to outliner. Therefore, (10) can be rewritten as,

$$q(s_i) \propto \exp\left(-\sqrt[3]{(s_i-\mu_i)^{\dagger}\Sigma_i^{-1}(s_i-\mu_i)}\right) \tag{15}$$

by applying (6) to (15), and by considering mean equal to zero and covariance matrix equal to identity matrix for (15), the score function becomes,

$$\varphi^{(k)}\left(\hat{s}_i(1),\ldots,\hat{s}_i(K)\right) = \frac{2\hat{s}_i(k)}{3\sqrt[3]{(\sum_{k=1}^{K}|\hat{s}_i(k)|^2)^2}} \tag{16}$$

In the presented research work, the performance of BSS is improved by a hybrid energy-driven model having multivariate generalized Gaussian and the original multivariate super-Gaussian source priors instead of identical source prior.

IV. PROPOSED HYBRID SOURCE PRIOR MODEL

In the proposed methodology, a multivariate generalized Gaussian source prior with heavy-tailed nature is used to exploit higher-order correlational information and other information is modeled by super-Gaussian source prior from the speech mixture. Therefore, the hybrid source prior model is mathematically written as,

$$q(s_i) = \phi f_{GGD} + (1-\phi) f_{SGD} \tag{17}$$

where f_{GGD} and f_{SGD} are the multivariate generalized Gaussian and multivariate super-Gaussian source priors respectively. $\phi \in [0, 1]$ is the weighting parameter, which determines weights of each source prior in the proposed hybrid model. Therefore, the multivariate hybrid score function is mathematically expressed as,

$$\varphi^{(k)}\left(\hat{s}_i(1),\ldots,\hat{s}_i(K)\right) = \phi\left(\frac{2\hat{s}_i(k)}{3\sqrt[3]{(\sum_{k=1}^{K}|\hat{s}_i(k)|^2)^2}}\right) + (1-\phi)\left(\frac{\hat{s}_i(k)}{\sqrt{\sum_{k=1}^{K}|\hat{s}_i(k)|^2}}\right) \tag{18}$$

The multivariate score function in (18) during the learning process preserves inter-frequency dependency in all frequency bins. The value of ϕ is frequency-dependent i.e., each frequency block has its weighting parameter $\phi(k)$. In this research work, the weighting parameter of each source prior distribution in the proposed hybrid model is adopted by the energy measure in the observed speech mixture. It is calculated as the normalized energy of the observed speech mixture for each frequency bin. The normalized energy of each frequency block is obtained by

$$E_b = \frac{1}{E_{tot}}\left(\sum_{k=f_b}^{l_b}||X_p(k)||^2\right) \tag{19}$$

where f_b and l_b is the first and last indices of the frequency block, respectively. $X_p(k)$ is the received mixture in the given frequency domain. E_b represents the energy of a particular frequency bin, and E_{tot} is the total received energy in the mixture. $||.||$ gives the respective Euclidean norm.

In this research work, more weight is given to generalizing Gaussian source prior than super-Gaussian prior if the frequency bin has high energy and vice versa. Therefore, it will better model the non-stationarity of the speech signal from the mixture.

The conventional IVA methods use only fixed source priors which can only either exploits second order statistics or high order statistics [8], [30], [33]. However, the proposed model can better exploit the statistical characteristics depending upon the received mixed speech signal. Therefore, the hybrid model enhances the performance of IVA algorithm and is more robust to the non-stationary environment.

V. RESULTS AND DISCUSSION

The performance evaluation of the proposed work is carried out based on proposed hybrid model using Matlab as a simulation tool. The algorithm is applied to the artificially generated mixture using a simulated room model.

A. OBJECTIVE EVALUATION

A pool of 10 speech signals are selected from TIMIT database comprising of 5 male and 5 female speakers [34]. Three different mixtures scenarios are evaluated for the proposed hybrid model and [8], [30], [33] i.e. male-male, male-female, and female-female speech mixtures. In each scenario, the results obtained from 5 speech mixtures for each window length, Fast Fourier Transform (FFT) frame length, and room reverberation (RT) parameter value is respectively averaged. Thus, a total of 15 speech mixtures with varying window length, 15 mixtures with FFT frames from 512 to 2048 frame length with fixed RT = 100ms, and 45 different mixed speech signals for different RT ranging from 40 to 200ms are used. These mixture signals are generated by simulated room model with room dimension 10 × 10×10m^3 [35]. The artificially generated mixture speech signals are fed into the proposed algorithm. The results are evaluated based on the obtained results from the proposed hybrid model and BSS methods [8], [30], and [33]. The RT is explicitly defined, whereas the speech signal sampling rate is considered to be 8 KHz. The BSS separation performance evaluation is based on signal to distortion ratio (SDR) in dB and ΔSDR is defined as the difference between desired SDR and speech mixture SDR. i.e. $\Delta\text{SDR} = \text{SDR}_{\text{desired}} - \text{SDR}_{\text{mixture}}$. The robustness of the proposed methodology is evaluated by varying different parameters used in these experiments

TABLE 1. Average results of different window length for male-male speech mixture.

Window Length	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
256	6.65	5.81	2.86	1.24	6.65	5.83	6.51	5.94
512	8.35	6.91	5.03	1.20	8.15	6.90	8.19	6.91
1024	9.51	6.47	5.66	0.90	9.14	6.48	9.16	6.51

TABLE 2. Average results of different FFT frame length for male-male speech mixture.

NFFT	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
512	8.16	6.78	5.86	1.73	8.16	6.79	8.02	6.76
1024	8.35	6.91	5.03	1.20	8.15	6.90	8.19	6.91
2048	8.35	6.91	4.51	0.98	8.34	6.90	8.18	6.91

TABLE 3. Average results of different RT for male-male speech mixture.

RT (ms)	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
40	16.80	13.55	9.79	2.74	16.75	13.60	16.87	13.72
60	13.19	10.57	8.19	2.13	13.15	10.58	13.18	10.66
80	10.34	8.34	6.30	1.57	10.30	8.44	10.33	8.49
100	8.35	6.91	5.03	1.20	8.15	6.90	8.19	6.91
120	6.97	5.85	4.01	0.85	6.97	5.85	6.84	5.84
140	5.95	5.06	3.37	0.63	5.96	5.07	5.92	5.10
160	5.24	4.55	2.98	0.59	5.25	4.57	5.24	4.54
180	4.68	4.10	2.70	0.46	4.69	4.13	4.66	4.08
200	3.43	2.95	1.45	0.31	1.65	1.33	4.18	3.69

such as, FFT frame length, window length and RT. During the experiment procedure, one parameter will be changed while the remaining parameters as describe above will remain unchanged. The convolutive mixed speech signal is comprised of two speech source signals. Three sets of experiments will be performed for male-male, male-female and female-female speech mixtures. Each set of experiment will be comprised of window length, FFT frame length and RT.

In the first set of experiments, 5 male speech source signals are obtained from the TIMIT database [34]. Different artificially mixed speech signals each containing two male source signals are generated by simulated room model having RT = 100ms [35]. The window length is varied from 256 window length to 1024 window length having a 75% overlap between adjacent windows. The FFT frame length is considered to be 1024. The remaining parameters are set as previously described. The results obtained from different speech mixtures are averaged and provided in Table 1. It shows that the highest SDR is achieved at 512 window length for the proposed hybrid model, [8] and [33] methodologies. Therefore, 512 window length is considered for the remaining set of experiments. The FFT frame length parameter is tested for better performance by varying it from 512 FFT frame length to 2048 frame length. Table 2 shows some improvement at 1024 FFT frame length than 512 and 2048 frame length for [8], [33] and proposed hybrid model. Thus, 1024 FFT frame length is assigned to the remaining experiments. In the last experiment of the set, different speech mixtures are generated from the male-male source speech signals by varying RT to access the performance of the proposed approach. The speech mixtures are generated by the simulated room model. It is observed in Table 3, that BSS performance of the proposed hybrid model, [8] and [33] methodologies in terms of SDR is comparable for male-male speech sources. However, the performance of [8], [33] and proposed hybrid model approach is better than Student's T source prior [30].

In the second set of experiments, 5 male and 5 female speech source signals are selected from TIMIT database [34]. Different speech mixture signals are generated by the same

TABLE 4. Average results of different window length for male-female speech mixture.

Window Length	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
256	5.01	2.44	6.38	3.19	6.01	2.87	5.04	2.37
512	8.80	3.06	10.86	2.35	8.82	2.92	8.90	3.29
1024	11.96	0.88	15.56	1.00	12.88	1.26	10.45	1.69

TABLE 5. Average results of different FFT frame length for male-female speech mixture.

NFFT	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
512	7.76	2.55	10.35	2.55	8.71	2.69	7.56	2.43
1024	8.80	3.06	10.86	2.53	8.82	2.92	8.90	3.29
2048	7.77	2.63	11.17	2.46	8.86	2.79	7.68	2.49

TABLE 6. Average results of different RT for male-female speech mixture.

RT (ms)	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
40	14.49	4.39	18.55	6.73	14.06	4.24	14.59	5.22
60	12.49	3.18	15.92	4.03	12.36	3.05	13.17	4.04
80	10.51	3.13	12.92	3.10	10.57	3.01	10.63	3.53
100	8.80	3.06	10.86	2.53	8.82	2.92	8.90	3.29
120	7.29	2.82	9.22	2.04	7.49	2.67	7.62	2.99
140	6.23	2.77	7.75	1.61	6.37	2.47	6.42	2.87
160	5.56	2.53	6.59	1.33	5.57	2.40	5.62	2.71
180	5.12	2.45	5.64	0.80	5.09	2.27	5.16	2.52
200	4.52	2.43	4.83	0.46	4.70	2.19	4.79	2.49

procedure for RT = 100ms. The window length is varied from 256 to 1024 sample length with 75% overlapping with neighboring windows. The FFT frame length is considered to be 1024 frame length. The trend in Table 4 shows magnificent improvement with 512 window length for the mentioned methodologies and the reason for its consideration for succeeding set of experiments. The FFT frame parameter is varied from 512 to 2048 frames for better performance assessment. Table 5 shows optimum performance at 1024 frame length for the proposed hybrid source prior model. So, 1024 frame length is fixed for the remaining experiment of the second set. In the last experiment of the set, speech mixtures are generated for different RT. The proposed hybrid source prior model provides improved system performance as compared to the multivariate Gaussian source prior [8], Multivariate Generalize Gaussian source prior [33]. However, Student's T source prior shows better performance for estimated S_1 than [8], [33] and proposed hybrid model, but the separation performance of [8], [33] and hybrid source prior shows better performance for estimated S_2. The results reflected in Table 6 shows that performance improvement reduces gradually by increasing RT, which is expected for high room reverberation.

We have selected 5 female speech signals from the TIMIT database for the last set of experiments. Different mixtures signal are generated by simulated room model for RT = 100ms. The window length is varied from 256 to 1024 samples. FFT frame length is considered to be 1024. Also, the results in Table 7 show improvement at 512 window length. As a result, 512 window length is fixed for the following experiments. The FFT frame is varied from 512 to 2048. The performance at 1024 shows improvement as shown in Table 8. Thus, it is considered for the remaining experiment. RT is varied having window length 512 and FFT frame 1024. Table 9 shows improvement for the proposed hybrid approach. The performance degrades slowly with the increment in RT but still shows enhancement as compared to multivariate Gaussian approach [8], [30], [33].

TABLE 7. Average results of different window length for female-female speech mixture.

Window Length	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
256	7.77	5.30	2.71	0.54	7.83	5.39	6.51	3.68
512	9.76	5.95	3.32	0.62	9.76	5.97	9.85	6.09
1024	10.43	5.58	3.87	0.24	10.53	5.68	10.71	5.82

TABLE 8. Average results of different FFT frame length for female-female speech mixture.

NFFT	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
512	9.75	5.88	3.84	1.18	9.75	5.92	9.56	5.87
1024	9.76	5.95	3.32	0.62	9.76	5.97	9.85	6.09
2048	9.73	5.95	3.12	0.35	9.75	5.96	9.52	5.85

TABLE 9. Average results of different RT for female-female speech mixture.

RT (ms)	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2	ΔSDR for Source-1	ΔSDR for Source-2
40	17.17	12.03	7.16	1.28	17.19	12.18	17.24	12.34
60	14.45	9.18	5.95	1.01	14.41	9.23	14.49	9.34
80	12.17	7.32	4.51	0.74	11.98	7.34	12.52	7.38
100	9.76	5.95	3.32	0.62	9.76	5.97	9.85	6.09
120	7.65	4.28	2.51	0.60	7.63	4.75	7.76	5.08
140	6.35	4.19	2.01	0.58	6.39	4.13	6.46	4.53
160	5.49	3.76	1.79	0.47	5.53	3.71	5.62	4.00
180	5.23	3.40	1.47	0.29	5.18	3.35	5.26	3.56
200	4.19	3.14	1.01	0.10	4.39	3.07	4.66	3.29

B. SUBJECTIVE LISTENING TESTS

The results of objective performance evaluation are cross verified by subjective listening tests. The subjective listening tests are conducted by 5 participants (3 male and 2 female) with normal hearing. Every listener is asked to mark a score ranging from integer 1(enhanced speech signal not audible) to 5 (enhanced speech signal clearly audible) for the estimated source signals separated from the mixture. Each participant is asked to listen to the original source signals and estimated speech signals. Note that, the participants have no prior knowledge of the BSS algorithms used to obtain the estimated speech signal.

The experiments are performed for male-male, male-female, and female-female scenarios. In these experiments, the speech mixtures used in the objective analysis are also used for the subjective listening tests. In the first experiment for male-male speech signals, the window length and FFT frame length are set to 512 and 1024 respectively. Different mixture signals are generated by the simulated room model for RT equal to 40, 80, 140, and 200ms. The score provided by the listening participants based on how clean the estimated speech signals are extracted from the mixtures. The estimated speech signals with less mixing signals have given the highest mean opinion score (MOS) and vice versa. The average results of MOS for male-male speech signals are reflected in Table 10. It is observed that the MOS score of the proposed hybrid model are comparable with multivariate Gaussian [8] as RT is increased. However, the results of the proposed model show improvement from Student's T [30] and Generalize Gaussian source prior models [33].

In the second experiment for the male-female scenario, previous values for window length and FFT frame length are considered. RT is considered to be 40, 80, 140, and 200ms. The MOS score results are shown in Table 11. This shows that the proposed hybrid model has improved MOS in comparison with [8] and [33], but Student's T source prior [29] shows

TABLE 10. Average MOS results obtained from subjective listening test with different RT for male-male speech mixture.

RT (ms)	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2
40	3.82	4.00	2.40	2.90	3.48	3.78	3.92	4.16
80	2.76	3.14	1.94	1.98	2.40	3.08	2.78	3.60
140	1.92	2.54	1.28	1.46	2.14	2.70	1.88	2.74
200	1.34	1.94	1.06	1.12	1.06	1.38	1.54	1.88

TABLE 11. Average MOS results obtained from subjective listening test with different RT for male-female speech mixture.

RT (ms)	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2
40	2.62	3.70	2.90	4.12	2.50	2.96	2.66	3.94
80	2.32	2.98	2.66	3.30	2.30	2.80	2.30	3.06
140	1.56	2.70	1.26	2.74	1.46	2.60	1.70	2.90
200	0.84	2.50	0.50	2.64	0.96	2.28	1.12	2.68

TABLE 12. Average MOS results obtained from subjective listening test with different RT for female-female speech mixture.

RT (ms)	Multivariate Gaussian Source Prior [8]		Student's T Source Prior [30]		Generalized Gaussian Source Prior [33]		Hybrid Source Prior Model	
	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2	MOS for Source-1	MOS for Source-2
40	2.84	2.82	1.50	1.46	2.80	2.98	3.08	3.16
80	2.46	2.69	1.48	1.34	2.38	2.76	2.82	2.86
140	2.24	2.44	1.28	1.02	2.34	2.44	2.42	2.52
200	1.56	2.00	1.00	0.66	1.80	1.88	1.93	2.12

improvement for RT equal to 40 and 80ms. However, as RT increases its results degrades significantly than the proposed hybrid model.

In the last experiment, the above-mentioned procedure is followed for female-female speech signals. The window length and FFT are considered to be the same as in previous listening testing experiments. Different mixture signals are generated for RT varying from 40 to 200ms. The average MOS results in Table 12 reflect an improvement of the proposed hybrid model in comparison with [8], [30] and [33].

C. TESTING IN THE PRESENCE OF NOISE

The experiments are also performed in noisy environment considering Additive White Gaussian Noise (AWGN). The parameters such as window length, FFT frame length, and RT are considered to be 512, 1024, and 100ms respectively. The results are obtained from 25 different speech mixtures randomly selected from the same pool of speech signals previously used. It is observed from Figure 1 and Figure 2 that the proposed source prior model shows improvement in the noisy environment from Gaussian source prior [8] for the estimated Source-1 and Source-2, respectively.

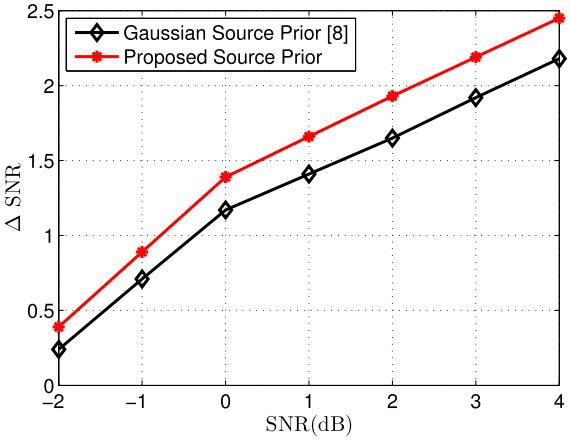

FIGURE 1. Gain in SNR for Source-1.

VI. PERFORMANCE EVALUATION

The proposed hybrid model separation performance is compared with a multivariate Gaussian source prior [8], Student's T source prior [30], and generalize Gaussian source prior [33]. In [8], the estimated source signals are extracted from the mixture signal by exploiting second-order statistics in the frequency domain. It is observed that the voice signals

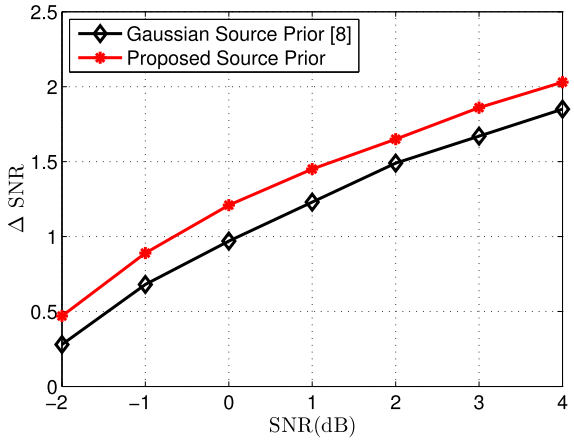

FIGURE 2. Gain in SNR for Source-2.

are non-stationary in nature and it contains low as well as high energy frequency components. Therefore, tracking of these non-stationary signals, multivariate Student's T [30], and generalize Gaussian distributions can be used as source prior [33]. These source priors can better model the non-stationary of the speech signals due to its heavy-tailed nature. In the proposed model, it proposed a BSS method which can model the low energy frequency components as well as high energy components. Therefore, a combination of two source priors is used to extract the low and high energy components from the mixture depending upon the energies in the frequency bins. Multivariate Gaussian source prior is used for a frequency bin with low energy components and multivariate generalize Gaussian source prior distribution is used for high energy components in a frequency bin.

The separation performance of the proposed approach is evaluated for three different mixtures scenarios i.e. male-male, male-female, and female-female speech mixtures. In each scenario, the results obtained from 5 speech mixtures for each window length, FFT, and RT parameter value is respectively averaged. Thus, a total of 15 speech mixtures with varying window length, 15 mixtures with FFT frames from 512 to 2048 frame length with fixed RT = 100ms, and 45 different mixed speech signals for different RT ranging from 40 to 200ms are used for the objective analysis using simulated room model. The same procedure is followed for [8], [30], [33].

In the first scenario, the proposed technique is compared with [8] for the male-male speech signal mixture. Table 1 and 2 show the average results in terms of SDR in dBs to select the window length and FFT frame length. It is reflected in Table 1 and 2 that the separation performance of window length and FFT frame length is better at 512 and 1024 respectively with fixed RT = 100ms. Therefore, these parameter values are considered for the remaining experiment for male-male speech mixtures. Table 3 represents the average results of varying RT from 40 to 200ms. It is observed in Table 3 that, the proposed hybrid source prior model achieved 0.06dB and 0.1dB enhancement for estimated source S_1 and S_2 in comparison with [8]. The proposed hybrid model is also evaluated with Student's T source prior distribution [30]. From Table 1 and 2 the window length and FFT frame length parameters are considered to be 512 and 1024 respectively. It can be seen in Table 3 that the proposed hybrid model shows a significant performance gain of 3.5dB for $\hat{S}_1$ and 5.8dB for $\hat{S}_2$ for different speech mixtures with RT varied from 40 to 200ms. The objective evaluation of the proposed methodology is also analyzed with [33]. The window length and FFT frames are considered as 512 and 1024 from Table 1 and 2. The evaluation of varying RT is displayed in Table 3, which shows an optimum gain 0.3dB for both estimated speech signals as compared to [33].

In the male-female scenario, the average results in terms of SDR are presented in Table 4, 5, and 6. The previous procedure is adopted in the male-male scenario is followed for male-female speech mixtures. The separation performance of the proposed hybrid model is evaluated with [8]. Both methodologies show enhancement at window length = 512 and FFT = 1024. Therefore, these parameters are considered for the succeeding RT experiment. The results of different RT varying from 40 to 200ms are demonstrated in Table 6. It is observed from Table 6 that, the proposed hybrid source prior improves its performance by 0.2dB for estimated $\hat{S}_1$ and 0.3dB for estimated $\hat{S}_2$ in comparison with [8]. In the comparison of the proposed model with [30] for male-female speech mixture signals. The parameters of window length and FFT are set to previous values as indicated in Table 4 and 5. Table 6 shows 1.7dB improvement of [30] for $\hat{S}_1$ from the proposed hybrid model. While for $\hat{S}_2$, the proposed approach outperforms [30] by 0.8dB. The comparison of proposed hybrid model with [33] is reflected in Table 4, 5 and 6. The window length and FFT frame length is set in accordance with Table 4 and 5. The average results are presented in Table 6 which indicates an overall improvement of in terms of SDR of 0.2dB and 0.5dB for $\hat{S}_1$ and $\hat{S}_2$ respectively as compared to [33].

The objective analysis for both female speech mixture results are presented in Table 7, 8, and 9. The window length and FFT frame length parameters are considered in accordance to Table 7 and 8, respectively. The number of speech mixtures are considered the same as in male-male and male-female scenarios. Table 9 shows the average results of different RT values from 40 to 200ms. The performance of proposed hybrid source prior is increased up to 0.2dB and 0.3dB for estimated source $\hat{S}_1$ and $\hat{S}_2$ respectively. With [30], the results are reflected in Table 7, 8, and 9. The parameters are set from Table 7 and 8. The RT results displayed in Table 9 shows magnificent improvement for the proposed model of 6dB for $\hat{S}_1$ and $\hat{S}_2$ from [30]. For the female-female speech mixture, the results of the proposed hybrid model are also compared with [33]. The parameters are adjusted from Table 7 and 8. The improvement of proposed methodology from [33] for different RT values is 0.2dB for both estimated speech signals.

A subjective listening testing performance evaluation tool is used for male-male, male-female, and female-female

speech mixture scenarios. The window length and FFT frame length parameters are considered as 512 and 1024 respectively. Also, 5 participants are selected to mark the score of estimated source signals. Each participant is asked to listen to the estimated speech signal obtained from the proposed and other BSS methods [8], [30], [33] for the above mentioned scenarios. The MOS score of the five participants is averaged for each RT value. The average MOS core presented in Table 10 shows an overall gain of the proposed hybrid model is 0.07dB for estimated $\hat{S}_1$ and 0.2dB for $\hat{S}_2$ comparable to [8]. Which indicates comparable gain for $\hat{S}_1$ and optimum gain for $\hat{S}_2$. In comparison with [30], the same procedural steps are taken for different mixtures with RT ranging from 40 to 200ms. The results in Table 10 show an enhancement of 0.9dB and 1.2dB for the hybrid model than [30] for the estimated source signals. The results of the proposed hybrid model are also compared with [33] for male-male speech mixtures. The overall average gain of the proposed method for $\hat{S}_1$ and $\hat{S}_2$ is 0.3dB and 0.4dB, respectively.

The subjective analysis for the male-female scenario is performed for [8] and the proposed method as shown in Table 11. It is observed that the proposed hybrid model performance is enhanced up to 0.1dB for $\hat{S}_1$ and 0.2dB for $\hat{S}_2$. The subjective listening test performs for [33] results are reflected in Table 11. The MOS results show a separation performance of 0.1dB for estimated $\hat{S}_1$ and 0.5dB for $\hat{S}_2$ for the hybrid model than [33].

The hybrid model is compared with [8], [30] and [33] for listening tests for female-female speech mixtures. It can be seen from Table 12 that the proposed hybrid source prior model improves its separation performance in terms of MOS up-to 0.3dB and 0.2dB for estimated $\hat{S}_1$ and $\hat{S}_2$ respectively from [8]. In comparison with [30], the average MOS results enhanced up to 1.2dB for $\hat{S}_1$ and 2dB for $\hat{S}_2$. The proposed model also shows improvement from [33] of 0.2dB for both estimated speech signals.

The comparative analysis reflects that the proposed hybrid model shows improvement for male-female and female-female scenarios than [8], [30] and [33]. In the case of male-male speech signals, the performance of [8] and the proposed model are comparable. While its separation performance is enhanced from [30] and [33].

The separation performance of the proposed model is compared with [8] in terms of computational complexity in time. The specifications of the computing system used to perform the experiments is intel(R) Core(TM) i3-4030U CPU with 1.90GHz processor and 8GB memory. It is observed that [8] method requires 1 minute and 7s to perform a single experiment while the proposed model performs the same experiment in 1 minute and 53s. the proposed method requires a little more time due to the energy calculation for every frequency bin.

VII. CONCLUSION AND FUTURE WORK

In this research work, a hybrid source prior is proposed which comprises multivariate Gaussian and generalized Gaussian source priors for blind separation of speech signals. The appropriate weights are being assigned between source priors in the hybrid model according to the underlying energy of the mixture speech signal. It effectively preserves the inter-frequency dependencies among different frequency bins as compared to the fixed source priors in [8], [30], [33]. The observed speech mixture contains both low as well as high energy components due to the randomness of speech mixture. The proposed hybrid model can adjust its weights of the source priors following the speech mixture energy in each frequency block to improve the separation process of the IVA approach. The simulation results clearly show significant improvement for speech mixture containing both low as well as high energy components from the conventional IVA model. In the future, the proposed approach can be extended for more complex scenarios such as reverberant speech mixture with a noisy environment. Furthermore, the hybrid model can be enhanced for the noisy reverberant environment to improve its robustness.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors therefore, acknowledge with thanks DSR for technical and financial support.

REFERENCES

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, Sep. 2005.
[2] Y. Xiang, I. Ubhayaratne, Z. Yang, B. Rolfe, and D. Peng, "Blind extraction of cyclostationary signal from convolutional mixtures," in *Proc. 9th IEEE Conf. Ind. Electron. Appl.*, Jun. 2014, pp. 857–861.
[3] B. Schwartz, S. Gannot, and E. A. P. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, Nov. 2017.
[4] S. Li and M. Stanaćević, "Gradient flow source localization in noisy and reverberant environments," in *Proc. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 257–260.
[5] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. Neural Netw. Signal Process. VI, IEEE Signal Process. Soc. Workshop*, Sep. 1996, pp. 423–432.
[6] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 45–48.
[7] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, Sep. 2019.
[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.
[9] R. M. Corey and A. C. Singer, "Speech separation using partially asynchronous microphone arrays without resampling," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 1–9.
[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.
[11] R. A. Khalil, S. M. Ashraf, T. Jan, A. Jehangir, and J. B. Khan, "Enhancement of speech signals using multiple statistical models," *Sindh Univ. Res. J.-SURJ, Sci. Ser.*, vol. 47, no. 3, pp. 519–522, 2015.
[12] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.
[13] L. Wang and A. Cavallaro, "Pseudo-determined blind source separation for ad-hoc microphone networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 5, pp. 981–994, May 2018.

[14] H. D. Hlynsson and L. Wiskott, "Learning gradient-based ICA by neurally estimating mutual information," in *Proc. Joint German/Austrian Conf. Artif. Intell.*, 2019, pp. 182–187.
[15] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *IEEE Trans. Neural Netw.*, vol. 19, no. 3, pp. 475–492, Mar. 2008.
[16] T. Jan, W. Wang, and D. Wang, "A multistage approach to blind separation of convolutive speech mixtures," *Speech Commun.*, vol. 53, no. 4, pp. 524–539, Apr. 2011.
[17] G. Fontgalland and P. I. L. Ferreira, "Combining antenna array elements by using ICA method for remote sensing of sources," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 234–237, 2017.
[18] S.-H. Hsu, T. R. Mullen, T.-P. Jung, and G. Cauwenberghs, "Real-time adaptive EEG source separation using online recursive independent component analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 3, pp. 309–319, Mar. 2016.
[19] C. Osterwise and S. L. Grant, "On over-determined frequency domain BSS," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 5, pp. 956–966, May 2014.
[20] D. Benzvi and A. Shafir, "An ICA algorithm for separation of convolutive mixture of periodic signals," in *Proc. IEEE Int. Conf. Sci. Electr. Eng. Isr. (ICSEE)*, Dec. 2018, pp. 1–5.
[21] M. P. Syskind, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2007. [Online]. Available: https://link.springer.com/book/10.1007/978-3-540-49127-9#about
[22] C. Wang, Y. Xu, M. Tang, and L. Wang, "Blind source separation based on variational Bayesian independent component analysis," in *Proc. IEEE 3rd Adv. Inf. Technol., Electron. Automat. Control Conf. (IAEAC)*, Oct. 2018, pp. 1614–1618.
[23] M. Anderson, T. Adali, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.
[24] T. Jan, H. Zafar, R. Khalil, and M. Ashraf, "A blind source separation approach based on IVA for convolutive speech mixtures," in *Proc. 8th Comput. Sci. Electron. Eng. (CEEC)*, Sep. 2016, pp. 140–145.
[25] S. Erateb, M. Naqvi, and J. Chambers, "Online IVA with adaptive learning for speech separation using various source priors," in *Proc. Sensor Signal Process. Defence Conf. (SSPD)*, Dec. 2017, pp. 1–5.
[26] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis, the Kotz distribution, and performance bounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3243–3247.
[27] Q. Li, S. M. Naqvi, J. Neasham, and J. Chambers, "Robust cooperative navigation for AUVs using the student's t distribution," in *Proc. Sensor Signal Process. Defence Conf. (SSPD)*, Dec. 2017, pp. 1–5.
[28] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, "Underdetermined source separation using time-frequency masks and an adaptive combined Gaussian-student's t probabilistic model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4187–4191.
[29] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex student's t-distribution for blind audio source separation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
[30] W. Rafique, S. M. Naqvi, P. J. B. Jackson, and J. A. Chambers, "IVA algorithms using a multivariate student's t source prior for speech source separation in real room environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 474–478.
[31] W. Rafique, S. Erateb, S. M. Naqvi, S. S. Dlay, and J. A. Chambers, "Independent vector analysis for source separation using an energy driven mixed student's t and super Gaussian source prior," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 858–862.
[32] S.-I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 757–763.
[33] Y. Liang, S. M. Naqvi, W. Wang, and J. A. Chambers, "Frequency domain blind source separation based on independent vector analysis with a multivariate generalized Gaussian source prior," in *Blind Source Separation*. Berlin, Germany: Springer, 2014, pp. 131–150.
[34] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. LDC93S1, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1
[35] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

JUNAID BAHADAR KHAN received the bachelor's degree in computer systems engineering from the COMSATS Institute of Information Technology, Islamabad, and the master's degree in electrical engineering from the Blekinge Institute of Technology, Sweden. He is currently pursuing the Ph.D. degree in electrical engineering with the Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.

His research interests include blind signal processing, blind reverberation time estimation, speech enhancement, compressed sensing, and non-negative matrix/tensor factorization for the blind source separation.

TARIQULLAH JAN received the bachelor's degree in electrical engineering from the University of Engineering and Technology Peshawar, Pakistan, in 2002, and the Ph.D. degree in the field of electronic engineering from the University of Surrey, U.K., in 2012.

He has been serving as an Associate Professor with the Department of Electrical Engineering, Faculty of Electrical and Computer Systems Engineering, University of Engineering and Technology Peshawar. His research interests include blind signal processing, machine learning, blind reverberation time estimation, speech enhancement, multimodal-based approaches for the blind source separation, compressed sensing, and non-negative matrix/tensor factorization for the blind source separation.

RUHUL AMIN KHALIL (Graduate Student Member, IEEE) received the bachelor's and master's degrees in electrical engineering from the Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering.

He has been serving as a Lecturer with the Department of Electrical Engineering, University of Engineering and Technology Peshawar. His research interests include audio signal processing and its applications, machine learning, the Internet of Things (IoT), routing, network traffic estimation, software defined networks, and underwater wireless communication.

ALI ALTALBE received the M.Sc. degree in information technology from Flinders University, Australia, and the Ph.D. degree in information technology from The University of Queensland, Australia.

He is currently working as an Assistant Professor with the Department of IT, King Abdulaziz University, Jeddah, Saudi Arabia. His current research interests include measuring information systems effectiveness, mobile services, e-learning, and human–computer interaction (HCI).