Learning Optimum Number of Bases for Indian Languages in Non-negative Matrix Factorization based Multilingual Speech Separation

Nandini C Nag[1], Milind S Shah[2]
Research Scholar, Electronics and Telecommunication Engineering Department[1]
Professor, Electronics and Telecommunication Engineering Department[2]
Fr. C. Rodrigues Institute of Technology, University of Mumbai[1, 2]
Vashi, Navi-Mumbai, 400703, India[1, 2]

***Abstract*—Non-negative matrix factorization-based audio source separation separating a target source has shown significant performance improvement when the spectral bases attained after factorization exhibits latent structures in the mixed audio signal comprising multiple speaker sources. If all the sources are known, the spectral bases may be inferred on priority by using a training process on the database of isolated sources. The number of bases inferred for a source should not include bases matching spectral patterns of the interfering sources in the audio mixture; otherwise, the estimated target source after separation will be incorporated with undesirable spectral patterns. It is difficult to distinguish and separate similar audio sources in an overlapped speech, leading to a complex speech processing task. Therefore, this research attempts to learn an optimum number of bases for Indian languages leading to successful separation of target source in multi-lingual multiple speaker speech mixtures using non-negative matrix factorization. The languages used for utterances are Hindi, Marathi, Gujarati, and Bengali. The speaker combinations used are female-female, male-male, and female-male. The optimum number of bases which was determined by evaluating improvement in the separation performance was found to be 40 for all the languages considered.**

***Keywords*—Indian languages; optimum number of bases; non-negative matrix factorization; speech separation**

I. Introduction

Separating audio source signals from a monaural recording is a complex problem. This problem is aggravated if the audio sources in the recording are overlapped with each other and are similar. A successful solution to these problems is compositional models, where the magnitude spectra of an audio signal can be decomposed into a linear combination of "spectral bases". Therefore, the bases for all the sources comprising the mixture combine linearly to constitute the magnitude spectra of the mixed audio signals. This leads to the fact that optimum estimation of the contribution made by the bases of a particular source to the mixed signal will help separate the said source.

Lee & Seung demonstrated non-negative matrix factorization (NMF) as a method that learns to represent a face as a linear combination of its "basis images". According to them, the basis images are local features corresponding to the parts of faces [1], [2]. The data matrix, in this case, is a non-negative image database which is NMF decomposed into two non-negative matrices, the part of the faces and their weights such that the original data matrix is approximated by their product.

The different domain using NMF expresses the columns of M (the data matrix) in terms of positively weighted sums of the columns of B (the parts or the basis vectors). Table I shows some examples of relations between the data matrix and the basis vectors or bases for some domains.

Apart from the above examples, this model became phenomenally successful as an audio source separation algorithm. It usually decomposes the spectrogram of an audio mixed-signal (M) into several "spectral bases" (B) and "temporal weights or activations" (A).

When the original data is corrupted, i.e., an audio signal is interfered with by simultaneous speakers and noise, NMF methods fail to learn an effective subspace or basis function matrix from the original data space or data matrix. In such cases, the basis functions matrix is populated with trained bases obtained by NMF decomposition of individual audio sources participating in the mixture. This basis functions matrix is then passed as a factor for data matrix (audio mixed signal) factorization, and only the activations matrix is updated. The estimated sources are obtained by multiplying the basis vectors with their corresponding activations. This procedure is called supervised source separation, as shown in Fig. 1, which provides improved separation performance. The limitation of such an approach is that it should know the sources prior to factorization.

TABLE I. Data Matrix and its Parts

Domain/ Application	M (data matrix)	B (parts or bases)
Computer Vision [1] [2]	Pictures of faces	Pictures of facial features
Document Clustering [3]	Documents	Base topics
Bioinformatics [4]	Spectra of chemical mixtures	Spectra of component molecules

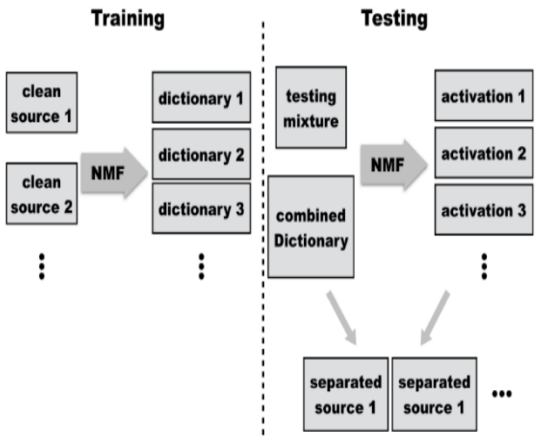

Fig. 1. Supervised Source Separation using NMF.

The separation problem is well studied for separating multiple speakers speaking the same language from a monoaural recording, but many multilingual overlapped speech recordings are not explored. A multilingual speech signal scenario is very usual in India, where 22 official languages are spoken. Any speech processing application addressing such speech mixtures as speech forensics or home assistant devices may find it challenging to recognize the desired speech leading to underperformance. Adding a speech separation module as a pre-processor to these applications in an Indian speech mixture scenario will help in improving the performance of segregating the desired speech. This leads to our motivation to further enhance the speech separation performance by identifying the bases obtained from individual sources, as discussed above in Fig. 1, which may better represent the mixed speech signal or the data matrix. The number of bases inferred for a speech source should not include bases matching spectral patterns of the interfering sources in the mixture; otherwise, the estimated target source after separation will be incorporated with undesirable spectral patterns.

Therefore, the objective is to learn the number of optimum bases representing individual Indian language speech sources to enhance the separation of one signal or all the participating signals from a multilingual, multiple-speaker speech signal comprising different Indian languages using NMF. The languages used are Hindi, Marathi, Gujarati, and Bengali. The evaluation metrics for separation performance were carried out by the "Blind Source Separation evaluation (BSS EVAL)" toolkit [5].

The organization of the paper is as follows: Related works are explained in Section 2, Methodology is elaborated in Section 3, Section 4 demonstrates the implementation, and Section 5 provides the results and discussion. The conclusion is given in Section 6.

II. Related Work

M.N Schmidt and R.K Olsson [6] proposed sparse NMF based source separation, which learns an over-complete set of non-negative basis vectors for each source. An over-complete set is a set where the number of bases is more than the spectral representation dimensions. According to the authors, better separation is achieved in separating individual audio sources from a mixture if each source is represented on an over-complete basis vector. The authors concluded that the dictionaries capture fundamental properties of speech; that is, the basis functions resemble phonemes. Convolutional NMF considers "spectro-temporal patterns" as bases instead of simple amplitude spectra in the paper [7]. This NMF variant extract cross-column patterns as single bases, therefore, capturing the temporal dependencies within bases.

Most of the previously discussed NMF-variants ignore individual signal phases and use the phase of the mixture signal while reconstructing the separated respective signals. This drawback of earlier NMF-variants introduced audible artifacts. Kameoka et al. in 2009 [8] presented an NMF-variant which allowed complex values and was given the name "complex non-negative matrix factorization". The authors proposed a mixing model called complex NMF established in the complex-spectrum domain. This paper aims to represent any observed complex spectrum where fewer active magnitude spectrum bases are paired with an arbitrary phase spectrum. King and Atlas in [9] named Complex NMF as "complex matrix factorization" (CMF). In this case, "each time-frequency point is multiplied with a phase term that allows each spectral base to assume the phase to fit the mixed-signal best", maintaining the non-negativity constraints of bases and activations.

Discriminative training of the NMF basis functions was introduced in [10], which generalized the model with separate analysis and reconstruction basis functions. Another research [11] selects active-set Newton algorithm (ASNA) for overcomplete NMF (over-complete set of basis functions), which outperforms other conventional source separation techniques. Simplex volume minimization [12] successfully estimates the source model, which learns an identifiable spectral basis. Working with dense basis matrix factors is allowed by these identifiability conditions. In addition, the basis matrix may have a full-column rank without any constraint imposed.

A pair of dictionaries was used for analysis and reconstruction in the paper [13]. It increases separation performance at low latencies, which is accomplished by utilizing shorter synthesis frames. According to the authors, if computational power is sufficiently available, this methodology may be applied to real-time applications. "Low-latency output allows a human listener to directly use the results of such a separation scheme without a perceptible delay". A binary subspace learning for the bases was proposed by [14]. Orthogonal NMF (ONMF) [15] adds orthogonality constraints to NMF in addition to the non-negativity constraint on one or both factors: the columns of B (bases) and the rows of A (activations) are required to be orthogonal. Newer variants of NMF [16] are being developed for hyperspectral and multispectral image fusion, which are yet to have been experimented with for audio source separation. Technologies other than NMF deliver competitive results in speech separation or enhancement, for example, deep learning neural networks (DNN) [17], but they are successful only with large training data. NMF is still suitable for a smaller dataset.

The number of bases retained during training differs for all the supervised speech separation discussed in existing research studies. Moreover, most of the research is based on a single language, primarily English or native languages. Therefore, studies on speech mixtures comprising different languages need attention. It is also crucial to identify the optimum number of bases or parts representing the latent structure of the data (mixed-signal) for successful separation of its comprising different language speech signals.

III. Methodology

The reason behind the successful separation of audio sources from a mixed signal using supervised NMF is the selection of an optimum set of basis vectors. Therefore, this section explains the methodology and the evaluation measures quantifying the separation performance. The performance is compared based on the metrics generated by BSS EVAL.

A. Non-negative Matrix Factorization

Positive Matrix Factorization was introduced by Paatero and Tapper in 1994, which was later coined as non-negative matrix factorization (NMF) [1]. Lee & Seung in 2001 [2] popularized NMF as a non-negative constrained algorithm capable of learning parts of faces from a facial image database favorably called image bases. The linear combination of these weighted parts constitutes each face. NMF decomposes the data (in this paper speech spectrogram) into basically two "non-negative components". The components are the "basis functions matrix," representing the spectrum of bases, and the "coefficient matrix," representing the activation coefficients of the bases in the data as in Fig. 2.

Recognizing NMF's capability, it was extended to several applications like "audio source separation," as explained in the Introduction. It separates the audio signal considered as target source from other interfering speakers or noise or music considered as maskers. It is possible to separate all the participating signals present in the audio mixture signal in some cases. The data representation of the mixed audio signal (M) is accomplished using spectrograms. The magnitude of the mixed audio signal (M) spectrogram is decomposed into basically two "non-negative components". The components are "basis functions matrix" (B) and "weight or activation or coefficient matrix" (A).

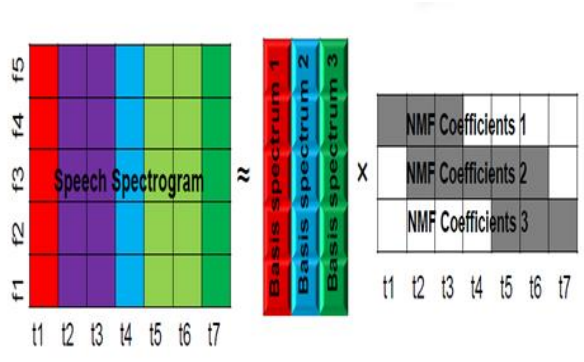

Fig. 2. A Speech Spectrogram is Factorized into bases and its Weights using NMF [25].

The interpretable factorization may be expressed as $M \approx BA$. BA is the matrix multiplication of B and A, where $M \in \mathbb{R}_{\geq 0}^{P \times Q}$ is subjected to the constraints of non-negativity $B \in \mathbb{R}_{\geq 0}^{P \times L}$ and $A \in \mathbb{R}_{\geq 0}^{L \times Q}$.

$P \in \mathbb{R}_{>0}$ is the number of the frequencies representing the spectrum of the mixed-signal M. $Q \in \mathbb{R}_{>0}$ is the time axis representing the mixed-signal M spectrogram. $L \in \mathbb{R}_{>0}$ is the number of the column basis vectors in B and activations row-wise in A. Cost functions along with multiplicative updates converge the non-negative factorization to a substantial approximation. For simplicity, M and BA are represented by X and Y for the following cost function expressions. X_{pq} and Y_{pq} are the elements (p=row, q=column) of the matrices X and Y, respectively.

The "Euclidean distance" (EUC) between X and Y [2] is given by:

$$\| X - Y \|^2 = \sum_{pq} \left(X_{pq} - Y_{pq}\right)^2 \quad (1)$$

The "Kullback-Leibler divergence" (KL) [2] is the cost function which leads to relative entropy when $\sum_{pq} X_{pq} = \sum_{pq} Y_{pq} = 1$.

$$div(X \| Y) = \sum_{pq}(X_{pq} \log \frac{X_{pq}}{Y_{pq}} - X_{pq} + Y_{pq}) \quad (2)$$

Another cost function given below is "Itakura-Saito (IS) divergence" [18]

$$div(X \| Y) = \sum_{pq}(\frac{X_{pq}}{Y_{pq}} - \log \frac{X_{pq}}{Y_{pq}} - 1) \quad (3)$$

Both the cost functions are non-increasing, which leads to minimization or convergence. The elements of B and A are initialized either randomly or using some pre-defined methodology with non-negative values. Convergence is achieved by executing the following multiplicative update theorems iteratively:

The EUC || M – BA || is updated by the following rules [2]:

$$A \leftarrow A \circ \frac{B^T}{B^T} \frac{M}{BA} \qquad B \leftarrow B \circ \frac{MA^T}{BAA^T} \quad (4)$$

The divergence div(M || BA) for KL uses the following [2] to update rules:

$$A \leftarrow A \circ \frac{B^T \frac{M}{BA}}{B^T.1} \qquad B \leftarrow B \circ \frac{\frac{M}{BA} A^T}{1.A^T} \quad (5)$$

For IS divergence, the multiple updates established by [18] is given by:

$$A \leftarrow A \circ \frac{B^T \frac{M}{(BA)^2}}{B^T.\frac{1}{BA}} \qquad B \leftarrow B \circ \frac{\frac{M}{(BA)^2} A^T}{\frac{1}{BA}.A^T} \quad (6)$$

B. Performance Measures

BSS Eval toolkit presents signal level metrics which evaluates the amount of speech enhancement or improvement and interference reduction. According to [5] the separated or estimated source $\hat{s}$ is expressed as a sum of the target source s_{target} and three types of error as follows:

$$\hat{s} = s_{target} + e_{interf} + e_{noise} + e_{artif} \quad (7)$$

"where s_{target} is part of the estimated source, which is the true source signal modified by a permissible distortion. The term e_{interf} is the error caused by interference from the unwanted sources. The sensor noise represented as the part of the estimated source is e_{noise}. The artifact error term, e_{artif}, is the part of the estimated source perceived as coming from other sounds, like forbidden disturbances and/or 'burbling' artifacts".

The ratios "source to distortion ratio" (SDR), "source to interference ratio" (SIR), and "source to artifact ratio" (SAR) over the audio signals are computed, which determines the relative value of each of these estimated target source and error terms given as follows:

$$\text{SDR: } = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \quad (8)$$

$$\text{SIR: } = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \quad (9)$$

$$\text{SAR: } = 10 \log_{10} \frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2} \quad (10)$$

The mixtures considered in the experiments conducted and mentioned in this paper are assumed to be noiseless. Therefore, only the SDR, SIR, and SAR performance measures are used throughout the experimentation. "SIR measures the quantum of the interfering sources present in the separated or estimated signal. The SAR measures the unwanted energy present in the signal that is not part of either the target or interfering audio signals. Combination of SIR and SAR into one measurement results in SDR".

IV. Implementation

Supervised NMF obtains basis vectors from individual speech sources participating in a mixed speech signal during the training phase. During the testing phase, these speech basis vectors from the training phase are used as the basis vectors matrix, which is one of the factors for the mixed speech signal factorization. The other factor is the activations matrix, which is updated, keeping the basis vectors matrix constant. The multiplication of basis vectors with the respective updated activations provides the separated signals. The experimental setup and evaluation methods engaged in this research are given below:

A. Experimental Setup

Synthetic mixtures of speech signals comprising different Indian languages are selected for the investigation, mainly taken from Hindi, Marathi, Gujarati, and Bengali speech audio databases. Bengali male (SLR37) [19], Marathi female (SLR64) [20], Gujarati male and female (SLR78) [20] multi-speaker speech databases are taken from openSLR (Open Speech and Language Resources) developed by Google. Hindi female, Bengali male, and Marathi male multi-speaker speech databases are taken from TTS voice data from IIIT Hyderabad [21]. Bengali female and male multi-speaker speech databases are also taken from the SHRUTI speech corpus developed by the Indian Institute of Technology; Kharagpur (IITKGP) distributed by the Society for Natural Language Technology Research [22].

The mixed speech signal was created by digitally combining male or female speech utterances of one language with male or female speech utterances of another language. For each language, the training data chosen was 60 utterances ranging from 3.00 to 5.00 seconds. The testing data selected was 5 utterances of similar duration different from training data. The testing data was augmented by combining one language utterance to all 5 utterances of another language, making it 25 utterances. One of the speech signals separated from the mixed speech signal is the target speech signal, and the other speech signal is called the interfering or the masker speech signal. The target speech signal to the masker speech signal is mixed with a target-to-masker ratio (TMRs) of 0 dB.

All the speech audio sources (WAV files) categorized for the training and testing phase were sampled at 16KHz. For the time-frequency (TF) representation, the short-term Fourier transform (STFT) was computed using 1024 points. A 32ms long with a 16ms overlap Hamming window was utilized for the same. The number of basis vectors experimented with for both the sources (all language combinations in this paper) was 40, 75, 100, and 150. The algorithm was executed at 500, 1000, and 1500 test iterations for each number of basis vectors chosen.

The different language speech combinations engaged are Marathi with Bengali, Marathi with Hindi, Hindi with Gujarati, Gujarati with Marathi, and Bengali with Hindi. The NMF cost function used was KL divergence for all the experiments. PYTHON programming language was used for the NMF algorithm with multiplicative updates. Parselmouth, PRAAT in PYTHON [23] was used for the spectrograms.

B. Evaluation

The source separation results were evaluated using the signal level metrics BSS_EVAL tool (SDR-source to distortion ratio, SIR-source to interference ratio, and SAR-source to artifact ratio), which quantifies the speech enhancement or interference mitigation.

V. Results and Discussion

This section analyses the separation performance results to learn the optimum number of basis vectors required for individual speech spoken in different Indian languages in a supervised audio source separation using NMF, which will subsequently help in successful speech separation from a mixed speech signal.

As mentioned in Implementation, the mixed speech signal comprises two speakers of the same or different genders (female-female, male-male, and female-male) speaking different Indian languages simultaneously. The language combinations are Hindi-Gujarati, Hindi-Bengali, Bengali-Marathi, Hindi-Marathi, Marathi-Gujarati. The speaker combinations are female-female, male-male, female-male.

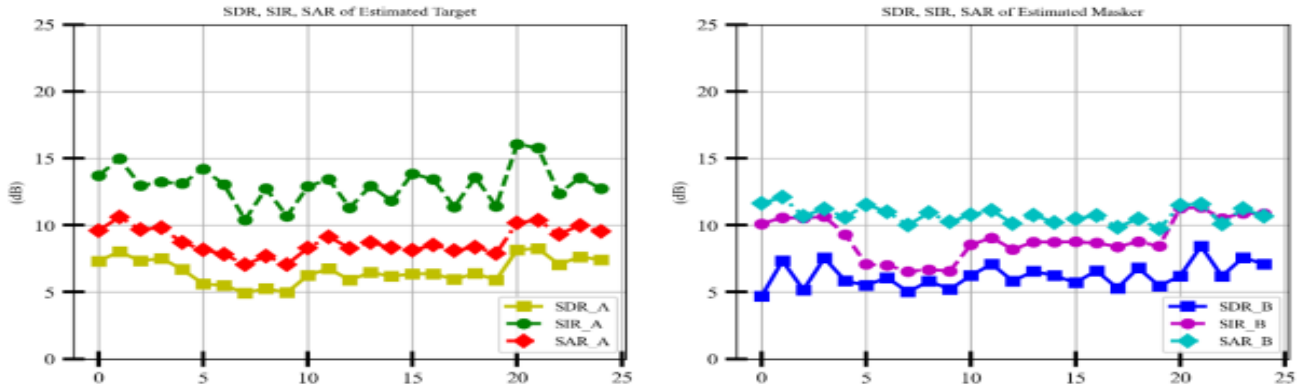

Fig. 3. Speech Separation Performance of Estimated Target (Hindi Female Speech) and Masker (Gujarati Male Speech) from 25 Testing Mixed Signals.

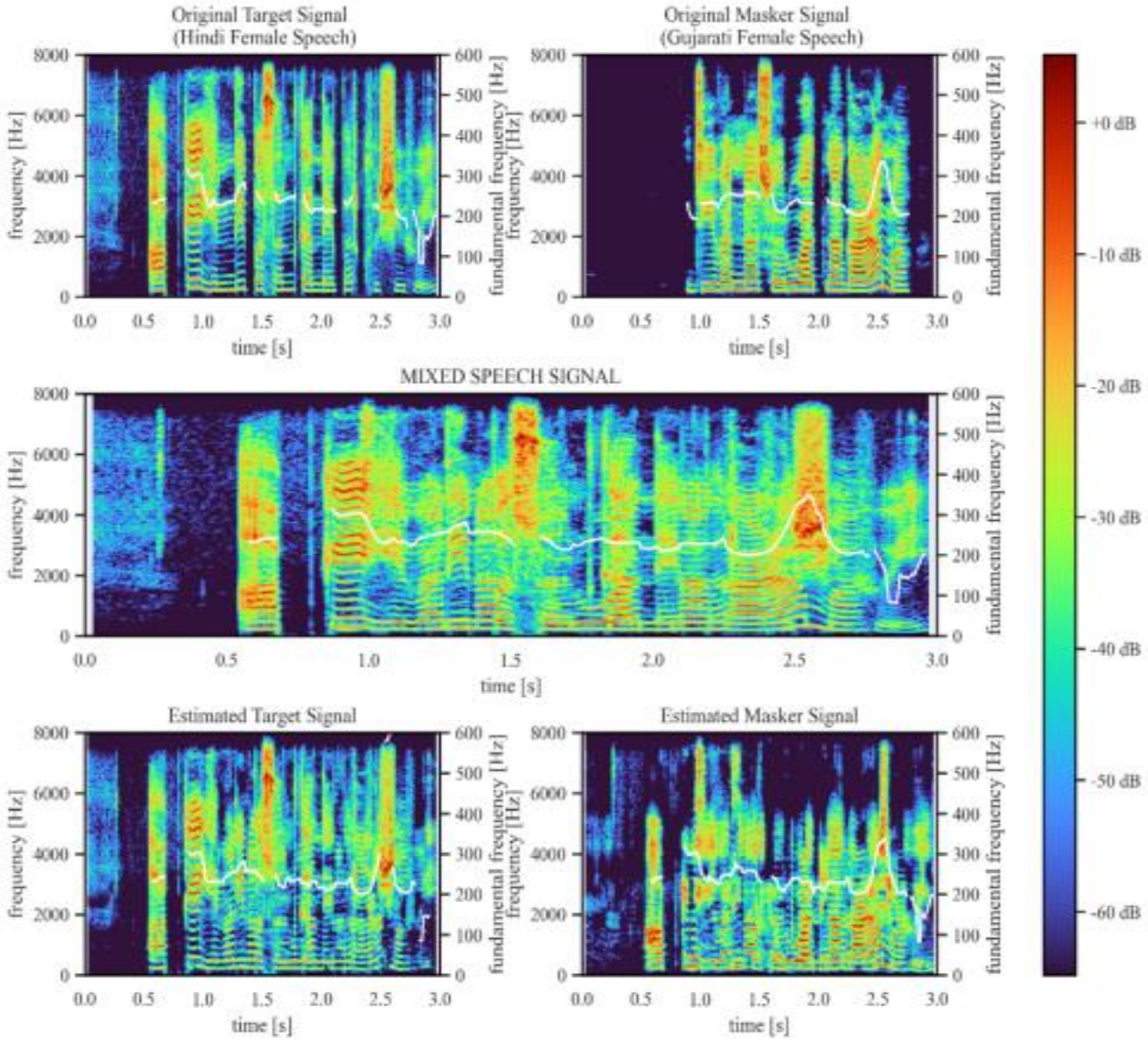

Fig. 4. Spectrograms of Original Target and Masker, Mixed Speech Signal and Estimated Target and Masker for Female-Female Combination.

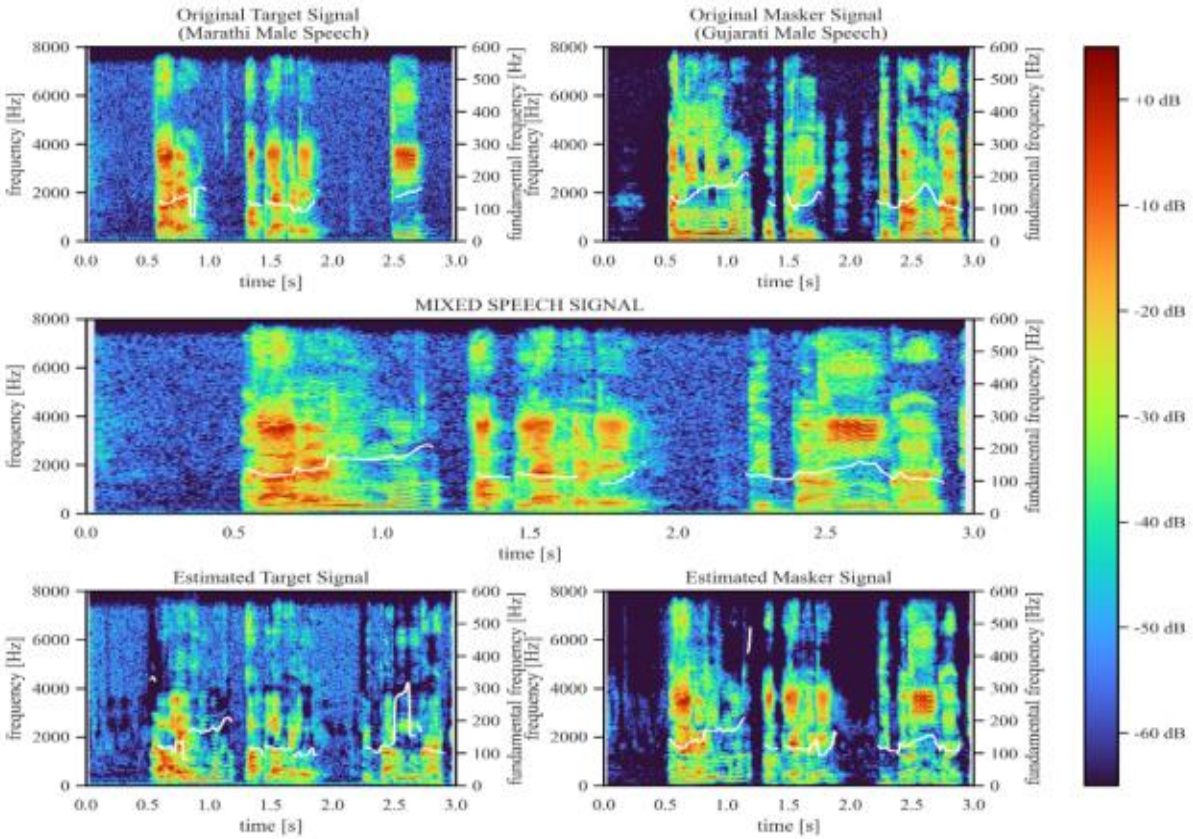

Fig. 5. Spectrograms of Original Target and Masker, Mixed Speech Signal and Estimated Target and Masker for Male-Male Combination Speaking Marathi and Gujarati, Respectively.

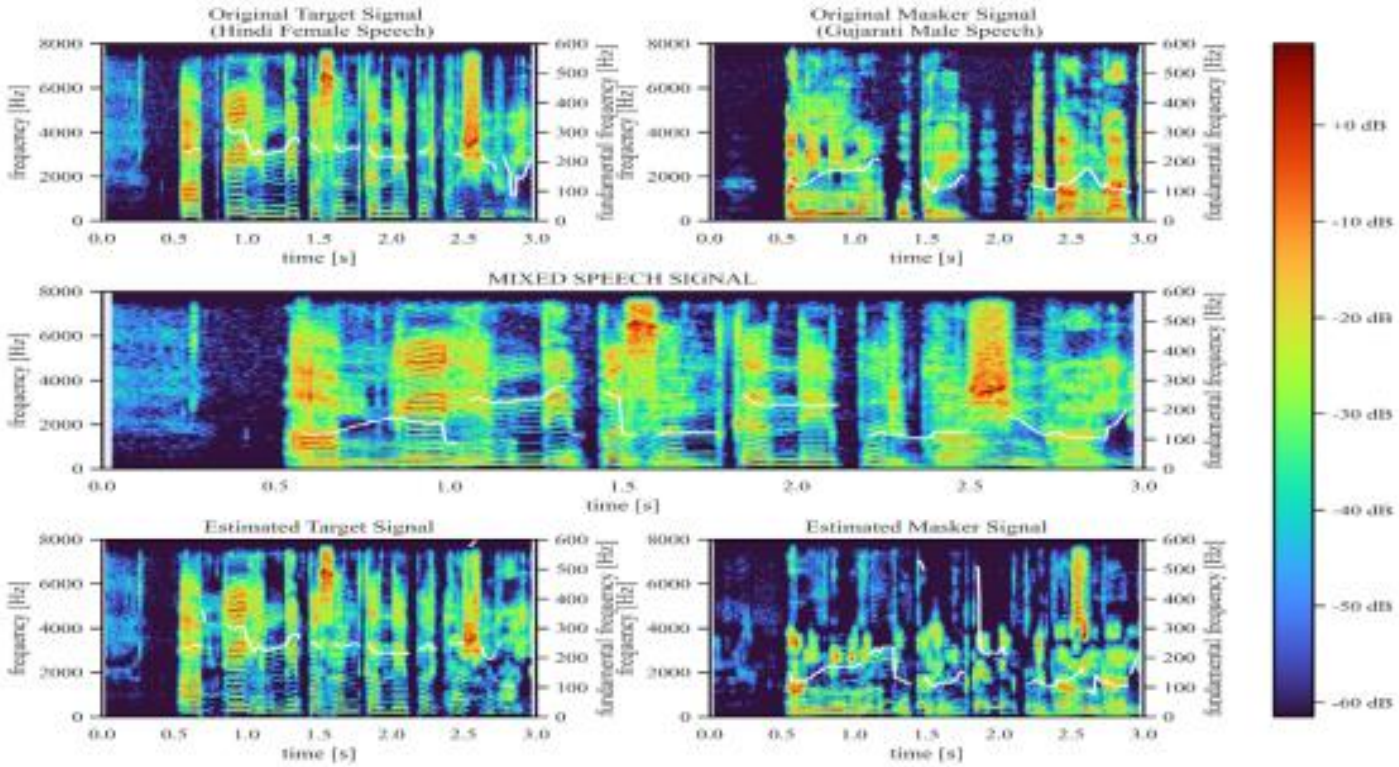

Fig. 6. Spectrograms of Original Target and Masker, Mixed Speech Signal and Estimated Target and Masker for Female-Male Combination Speaking Hindi and Gujarati, respectively.

For each of the combination BSS EVAL parameters, SDR, SIR, and SAR are used to quantify separation quality. As interpreted from the Implementation section, we have 25 mixed utterances in the testing phase; the evaluation parameters were computed for all the separated utterances. The same is displayed in Fig. 3 for one such combination. As evident from the figure, the deviation for the parameter values from the mean value is significantly less, within the range of 2 to 3 dB; therefore, the mean value is considered for this and other language-gender combinations.

Spectrograms are used to display the speech parameters. Fig. 4, Fig. 5, and Fig. 6 show the spectrograms of original target and masker, mixed speech signal, and estimated target and masker for female-female, male-male, and female-male, respectively, which displays speech separation. The pitch is highlighted in the spectrograms to show the interference, which is not a very significant presence in the separated target speech due to the masker speech and vice-versa.

NMF on the mixed speech signal for each pair of language-gender combinations was experimented with for basis vectors 40, 75, 100, and 150 to assess the optimum number of bases. Each set of basis vectors are obtained in the training phase with an updated iteration fixed at 1500. Each set of basis vectors was experimented with three different update iterations in the testing phase, namely 500, 1000, and 1500. For the remaining section, the number of iterations mentioned will be indicative of the testing phase.

The language combinations for female-female speech separation selected are Hindi-Bengali, Hindi-Gujarati, and Hindi-Marathi. The BSS EVAL parameters for one of the combinations (Hindi-Bengali) target speech and masker speech are tabulated in Table II.

TABLE II. BSS EVAL PARAMETERS OF HINDI FEMALE (TARGET) AND BENGALI FEMALE (MASKER) SEPARATED FROM A MIXED SIGNAL

Iteration	**Bases**	**Target**			**Masker**		
500		SDR	SIR	SAR	SDR	SIR	SAR
	40	3.11	4.67	5.60	2.03	3.36	7.64
	75	2.88	4.55	5.42	2.12	2.50	9.51
	100	2.75	2.65	6.65	1.82	1.69	9.21
	150	3.07	3.14	7.49	1.94	1.88	10.49
1000							
	40	2.44	3.04	5.34	1.82	2.00	7.36
	75	2.78	3.92	5.67	2.01	2.22	9.09
	100	2.78	2.77	6.33	1.85	1.90	8.84
	150	2.79	2.65	6.77	1.89	1.66	9.69
1500							
	40	3.62	6.29	5.49	2.24	4.52	8.01
	75	2.87	3.82	5.58	1.96	2.52	8.34
	100	2.92	4.04	5.56	2.03	2.46	8.95
	150	2.94	3.41	6.17	1.97	1.98	9.78

TABLE III. BSS EVAL PARAMETERS OF MARATHI MALE (TARGET) AND GUJARATI MALE (MASKER) SEPARATED FROM A MIXED SIGNAL

Iteration	**Bases**	**Target**			**Masker**		
500		SDR	SIR	SAR	SDR	SIR	SAR
	40	2.43	2.59	6.02	2.46	1.81	7.01
	75	3.61	4.45	7.39	3.42	3.49	9.35
	100	3.04	2.57	7.69	2.86	1.94	9.11
	150	3.29	2.57	9.00	2.94	2.25	9.38
1000							
	40	2.07	1.74	6.47	2.16	1.00	5.91
	75	2.24	1.22	7.49	2.16	0.59	6.73
	100	2.76	2.08	7.09	2.67	1.57	7.88
	150	3.09	2.44	7.99	2.89	1.96	9.31
1500							
	40	2.33	2.39	6.59	2.31	1.82	5.69
	75	2.63	2.45	6.25	2.62	1.80	7.38
	100	2.73	2.34	7.16	2.65	1.38	8.13
	150	3.40	3.14	8.01	3.12	2.81	8.76

The language combinations for male-male speech separation selected are Marathi-Gujarati, Bengali-Gujarati, and Bengali-Marathi. Table III tabulates the BSS EVAL parameters for one of the combinations (Marathi-Gujarati) target speech and masker speech. The language combinations for female-male speech separation selected are Hindi-Bengali, Hindi-Gujarati, Bengali-Marathi, Bengali-Gujarati, and Hindi-Marathi. One of the combinations (Hindi-Gujarati) target speech and masker speech BSS EVAL parameters are tabulated in Table IV.

Comparing the spectrograms of Fig. 4, 5, and 6 reveals that the estimated target and the masker have interferences from the other speaker's utterance, but they are insignificant. Careful observations show the quantum of interference is more in female-female and male-male than female-male combination. It is well understood as NMF is based on spectral bases. More are the similarity in spectral bases of the source speakers; less is the separation performance as it is difficult to distinguish similar frequencies. Therefore, the separation in the female-male combination is better as their speech fundamental frequencies are at different levels (i.e., male: 80-180 Hz and female: 160-250 Hz).

Now let us consider BSS EVAL parameter SDR for an estimated target separated from a different language female-male combination mixed speech signal. The results are shown as 3D bar plots in Fig. 7, 8, and 9, which show the comparison between the SDR of an estimated target separated from a Hindi-Bengali, Hindi-Gujarati, and Bengali-Marathi mixed signal for update iteration 500, 1000, and 1500, respectively. The SIR and SAR values are discussed from the tables mentioned above.

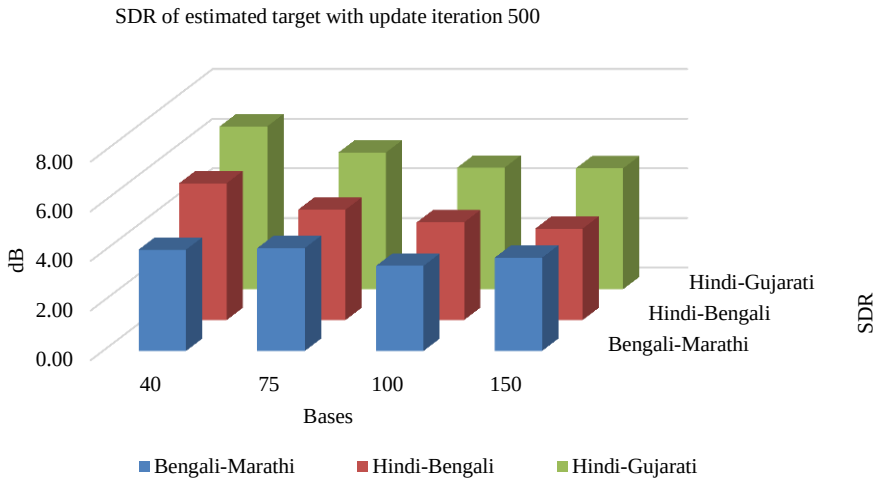

Fig. 7. Comparison of SDR of Estimated Target for Hindi Female (Target)-Bengali Male (Masker), Hindi Female (Target)-Gujarati Male (Masker) and Bengali Female (Target)-Marathi Male (Masker) Separated from a Mixed Signal with Update Iteration 500.

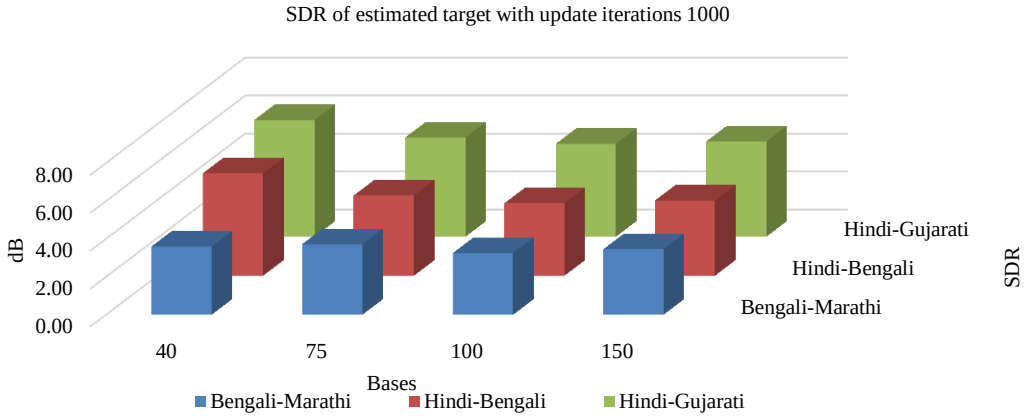

Fig. 8. Comparison of SDR of Estimated Target for Hindi Female (Target)-Bengali Male (Masker), Hindi Female (Target)-Gujarati Male (Masker) and Bengali Female (Target)-Marathi Male (Masker) Separated from a Mixed Signal with Update Iteration 1000.

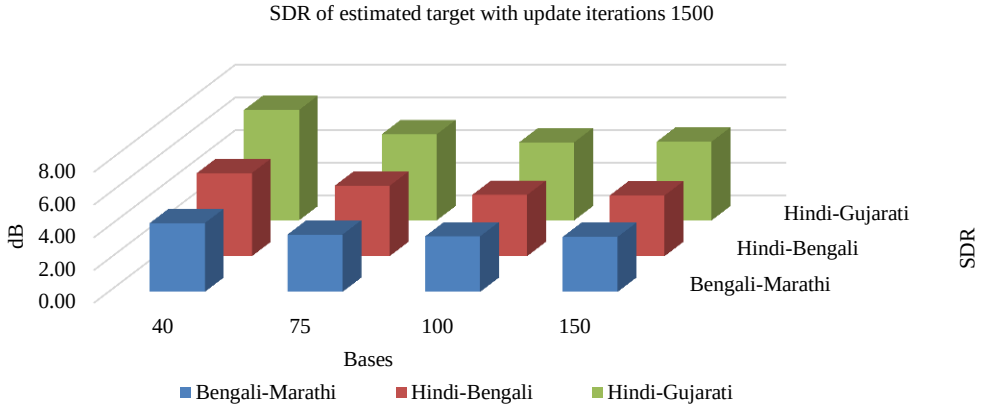

Fig. 9. Comparison of SDR of Estimated Target for Hindi Female (Target)-Bengali Male (Masker), Hindi Female (Target)-Gujarati Male (Masker) and Bengali Female (Target)-Marathi Male (Masker) Separated from a Mixed Signal with Update Iteration 1500.

From the above-mentioned plots, it is observed that the language combination Hindi-Gujarati (female-male) showcases the best result with SDR (estimated or separated target speech) almost nearing 7 dB for update iteration of 500 and 1500 in case of 40 basis vectors. It is also evident from Table IV. Another observation is for most combinations; the SDR is highest for 40 bases though the actual value between the language combination differs by 0.7 dB to 1.5 dB. Though the Bengali-Marathi (male-male) combination shows the lowest results, it shows the highest value of 4.15 dB and 3.72 dB in the case of 75 basis vectors for update iteration 500 and 1000, respectively.

Table III shows the language combination Marathi-Gujarati (male-male) exhibits higher SDR results with 40 basis vectors for 500 iterations followed by 150 bases for 1500 iterations applicable for both the target and the masker speech sources. Table II shows Hindi-Bengali (female-female) language combination for mixed speech signals. It is noticed that SDR values of both the target and the masker speech sources are higher for 500 and 1500 iterations with 40 basis vectors. The SIR result reflects the same as SDR. However, SAR results show higher results with 150 basis vectors for all the iterations, with the only exception in the female-male case where it displays higher results with 40 basis vectors for 1500 iteration. For all the combinations, the effects of update iterations 500 and 1500 are better than 1000. It is noticed that almost all the results suggest 40 basis vectors to be the optimum number after comparing the performance with respect to bases.

As mentioned above, the supervised separation performance of NMF, which is known for its reduced dimensionality depends on the bases representing the latent structures in the mixed speech signal; the objective of this study was to learn the optimum number of bases representing Indian language speech sources in a mixed signal.

TABLE IV. BSS EVAL PARAMETERS OF HINDI FEMALE (TARGET) AND GUJARATI MALE (MASKER) SEPARATED FROM A MIXED SIGNAL

Iteration	**Bases**	**Target**			**Masker**		
500		SDR	SIR	SAR	SDR	SIR	SAR
	40	6.56	13.02	8.77	6.21	9.04	10.77
	75	5.51	11.40	8.05	5.56	6.87	11.34
	100	4.90	9.97	7.80	5.09	5.80	11.74
	150	4.87	9.72	8.28	5.01	5.45	12.07
1000							
	40	6.16	12.95	8.07	6.00	8.56	10.36
	75	5.23	11.89	7.09	5.42	6.72	10.72
	100	4.90	10.78	7.19	5.16	6.04	11.14
	150	5.02	9.08	8.16	5.07	5.91	11.73
1500							
	40	6.75	12.91	9.08	6.31	9.35	10.94
	75	5.26	10.82	7.40	5.35	6.85	10.70
	100	4.78	10.49	7.06	5.08	5.80	11.23
	150	4.81	9.00	7.76	4.93	5.65	11.42

There is no fixed directive to identify the number of bases; the same was learned by utilizing a different number of basis vectors. Each set was used for a different number of iterations in the testing phase. The separation performance is at its best when the bases resemble phonemes or speech sounds of the language. From the literature study, it is known that the languages Hindi, Bengali, Marathi, and Gujarati are Indo-Aryan languages, and their phoneme ranges from 37 (Bengali) to 52 (Marathi) [24]. It is, therefore, understandable that the optimum number of spectral bases required for the individual speech source signal of different Indian languages emerging is 40, after comparing all the speech separation results delivered by NMF.

VI. CONCLUSION

Supervised speech separation of a desired or target speech source from a multi-lingual two-speaker speech mixture is considered, which is very relevant to an Indian scenario as India is a country with a vast population speaking different languages. For successful separation proper set of bases needs to be inferred from the participating speech sources in the mixed signal, i.e., bases matching spectral patterns of the interfering sources in the mixture should not be included as the estimated target source after separation may be incorporated with undesirable spectral patterns. Therefore, this research attempts to learn an optimum number of bases for Indian languages using non-negative matrix factorization. Hindi, Marathi, Gujarati, and Bengali Indo-Aryan languages are used for utterances. The speaker combinations used are female-female, male-male, and female-male.

The optimum number of bases determined by evaluating the separation performance for the individual speech source signal of different languages is observed as 40. This number is nearly like the phoneme sets of the languages engaged, which signifies that separation performance is better when the bases resemble phonemes or the speech language sounds. Though the number of bases is similar for all the languages, the separation performance parameter SDR shows different values for different language combinations. This difference in SDR values needs more insight into language correlation.

A pre-processor separating different language speech sources may be added to several speech processing applications, for example, audio or speech forensics, home assistant devices operating in Indian scenarios, thereby enhancing the applications' performance. The research can be continued for other Dravidian Indian languages, NMF variants and DNN may be utilized depending on the availability of the training dataset.

REFERENCES

[1] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788-791, 1999.

[2] D.D Lee and H.S. Seung, "Algorithms for nonnegative matrix factorization," Neural Inf. Process. Syst., vol. 13, pp. 556-562, 2001.

[3] Wei Xu, Xin Liu, Yihong Gong, "Document Clustering Based On Non-negative Matrix Factorization," SIGIR Forum, 2003.

[4] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov, "Metagenes and Molecular Pattern Discovery using Matrix Factorization," Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 12, pp. 4164-4169, 2004.

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

[6] M.N Schmidt and R.K Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in Proc Interspeech, 2006.

[7] Smaradis, P., "Convolutive Speech Bases and Their Application to Supervised Speech Separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 1-12, 2007.

[8] H. Kameoka et al., "Complex NMF:A new sparse representation for acoustic signals," in Proc. ICASSP, 2009.

[9] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in Proc. ICASSP, Dallas, TX, 2010.

[10] Felix Weninger, Jonathan Le Roux, John R Hershey, Shinji Watanabe, "Discriminative NMF and its application to single-channel source separation," Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[11] T. Virtanen, J. Gemmeke, B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11, pp. 2277-2289, 2013.

[12] Jianyu Wang, Shanzheng Guan, Shupei Liu, Xiao-Lei Zhang, "Minimum-volume Multichannel Nonnegative Matrix Factorization For Blind Source Separation," arXiv, 2021.

[13] Tom Barker, Tuomas Virtanen, Niels Henrik Pontoppidan, "Low-Latency Sound-Source-Separation Using Non-Negative Matrix Factorisation With Coupled Analysis And Synthesis Dictionaries," in ICASSP, 2015.

[14] Xiangguang Dai et al., "Robust semi-supervised non-negative matrix factorization for binary subspace learning," Complex & Intelligent Systems, 2021.

[15] Moses Charikar, Lunjia Hu, "Approximation Algorithms for Orthogonal Non-negative Matrix Factorization," in AISTATS 2021, 2021.

[16] Priya K, Dr. Rajkumar K K, "Multiplicative Iterative Nonlinear Constrained Coupled Non-negative Matrix Factorization (MINC-CNMF) for Hyperspectral and Multispectral Image Fusion," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.

[17] Norezmi Jamal, N. Fuad, MNAH. Sha'abani, "A Hybrid Approach for Single Channel Speech Enhancement using Deep Neural Network and Harmonic Regeneration Noise Reduction," International Journal of Advanced Computer Science and Applications, vol. 11, no. 10, 2020.

[18] Févotte et al., "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Computation, pp. 793-830, 2009.

[19] Keshan Sodimana et al., "A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese," in Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), Gurugram, India, 2018, pp. 66--70.

[20] He, Fei et al., "Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems," in Proceedings of The 12th Language Resources and Evaluation Conference (LREC), Marseille, France, European Language Resources Association (ELRA), 2020, pp. 6494--6503.

[21] Prahallad Kishore, Kumar E, Keri Venkatesh, Suyambu Rajendran, Black Alan, "The IIIT-H Indic Speech Databases," 2012.

[22] Biswajit Das, Sandipan Mandal and Pabitra Mitra, "Bengali speech corpus for continuous automatic speech recognition system," in COCOSDA, Taiwan, 2011.

[23] Yannick Jadoul, Bill Thompson and Bart de Boer, "Introducing Parselmouth: A Python interface to Praat," Journal of Phonetics, vol. 71, pp. 1-15, 2018.

[24] George Cardona and Dhanesh Jain, THE INDO-ARYAN LANGUAGES, London and New York: Taylor and Francis, 2007.

[25] Mohammadiha, Nasser, "Speech Enhancement Using Nonnegative Matrix Factorization," Department of EE, KTH Royal Institute of Technology, Stockholm, Sweden, 2013.