International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

https://dx.doi.org/10.12785/ijcds/110103

Non-negative Matrix Factorization on a Multi-lingual Overlapped Speech Signal: A Signal and Perception Level Analysis

Nandini C Nag[1] and Milind S Shah[2]

[1,2]*Electronics and Telecommunication Engineering Department, Fr. C. Rodrigues Institute of Technology, University of Mumbai, Sector 9A, Vashi, Navi Mumbai 400703, India*

Received 1 Jan. 2021, Revised 27 Jun. 2021, Accepted 8 Jul. 2021, Published 9 Jan. 2022

Abstract: A complex acoustic scenario comprising overlapping speeches from multiple speakers in the presence of noise renders speech recognition perform poorly in hands-free devices. This scenario turns out to be more complex in India, a country where 96.71% of the population speaks one of the 22 scheduled languages. Therefore, an audio source separation algorithm that mitigates the interference from other speakers and effectively enhances the articulacy and quality of source speech may be added as a pre-processor in speech recognition systems. This research, therefore, investigates the non-negative matrix factorization (NMF) algorithm's effectiveness for the separation of source in an overlapping multi-lingual multi-dialect single-channel speech mixture scenario, an inherent characteristic of a cocktail party problem in India. The objective is to analyze the signal level metrics and perception level metrics of a speech source-separated from a multi-lingual overlapped speech signal. The languages used for the same are English and two Indo-Aryan languages, Marathi and Bengali. One of the experimental results demonstrated that the source to distortion ratio (SDR) of separated target source from English-Bengali and English-Marathi speech mixture is 0.4 and 1.3 dB higher than English-English speech mixed signals, respectively. Therefore, the experiments highlight an improvement in separating sources from mixed speech signals with different language combinations than the same language.

Keywords: Audio source separation, Cocktail party problem, Multi-lingual scenario, Non-negative matrix factorization

1. Introduction

The modern era of communication involves the wireless exchange of messages among humans and devices. As we know, the simplest means of communication is speech, and due to this, we have seen enormous growth in speech-enabled hands-free devices, smart speakers, smart television (TV), voice assistant devices, etc. For seamless communication among these devices, speech recognition plays an important role. Speech recognition fails to acknowledge the voice commands or the speech signals of interest when interfered with either by background noise or multiple speakers: a complex acoustic scenario called a cocktail-party problem [1]. Listening to such a scene, a human brain can segregate a single speaker, music, noise, or any other audio source by setting up different aural objects or sound streams arriving at the ears [2]. The challenging task of sound stream separation has been implemented by developing audio source separation (ASS) algorithms. These algorithms deal with several distinguishing characteristics of source separation, like the number of sensors observing the audio source mixture signals. If the number of sensors or microphones is less than the number of sources, it is termed an "underdetermined system". The worst case is if the mixture signals are observed by only one sensor, a condition termed a "single-channel source separation (SCSS)" problem.

It is evident that the underdetermined system problems are ill-posed and may require a priori information of the sources for effective separation performance. However, SCSS may be helpful in applications where installing a microphone array (microphones positioned at different locations) is impossible due to spatial problems or power constraints, such as for in-the-ear hearing aids. The cost of the hardware also increases with the number of microphones, which may be another constraint. "Multi-channel source separation" (MCSS) is when more sensors than the number of sources observe the mixture signal. Therefore, they are aware of target and interference signal sources' spatial locations, which provide information like the direction of arrival (DOA), and separation is done by dereverberation and beamforming [3].

However, in both the source separation problems, they

E-mail address: nandini.nag@fcrit.onmicrosoft.com, milind.shah@fcrit.ac.in

are subject to different input scenarios, such as mixtures of various audio sources and their mixing conditions. It is difficult to distinguish similar audio sources such as overlapped speech, making it a complicated audio signal processing task. The source separation task may further be classified based on the information available about the sources. If all the sources are known, the ASS algorithm may infer the source estimate parameters or bases beforehand by using a training process on a database of isolated sources. This procedure is termed “supervised source separation”. Similarly, a source separation task is unsupervised if no training processes are used to infer source estimate parameters beforehand.

Therefore, the source separation task using an ASS algorithm may segregate a target speech signal or all signals involved in a mixture with or without noise [4]. It may also separate a speech signal from a combination of speech and music [5] or separate a singing voice from a musical composition [6]. The signals in some signal processing tasks are split based on “correlation and homogeneity properties” [7]. Another scenario may be to decompose a sound signal into four frequency bands [8].

Some of the well-researched ASS algorithms on different scenarios are Computational Auditory Scene Analysis (CASA) [9], [10], Blind Source Separation (BSS) [11], [12], Non-negative Matrix Factorization (NMF) [13], [14], Sinusoidal Modeling [15], Robust Principal Component Analysis [6] and Independent Component Analysis (ICA) [16]. One of the ASS techniques that went on to become a popular technology is NMF. Initially, it was demonstrated in a seminal paper [13] that NMF bounded by “non-negativity constraints” can learn parts of the face from face images. It is also capable of learning semantic features of the text in documents. This led to applications like topic modelling [17], [18]. NMF was explored further in audio source separation (ASS) applications.

Recent technique Deep Neural Networks (DNNs) are successful in several applications like speech and emotion recognition [19], music composition and classification [20], [21]. Utilizing DNNs for the separation of mixed audio signals started in 2014. Although the use of DNNs showed considerable improvement in separating speech and noise [22], [23], its performance on actual recorded noisy speech is not satisfactory. Moreover, it is not applicable in a cocktail party scenario, comprising overlapped speech [24]. They also point out that there is immense scope for improving audio source separation in overlapping speech scenarios. DNN, though it shows promising results in separation performance, is characterized by high computational complexity and suffers degraded performance on problems with limited training data or small data sets [25]. NMF, on the other hand, is still prevalent for separation with limited training datasets. Moreover, newer models and algorithms are still being proposed for NMF by researchers [26], [27]. In this study, therefore, NMF is considered for the ASS algorithm.

The need for an audio source separation (ASS) system comes from its usefulness to many applications, for example, interference suppression for mobile phones, hearing aids, home assistant devices, smart speakers, and TV. Despite the advancement in speech recognition capabilities, these intelligent devices fail to understand the voice commands overlapped with speech (multiple speakers speaking simultaneously) or noise. ASS may be used as a preprocessor to such speech recognition module, enhancing its recognition capabilities and reducing processing overload.

The separation performance not only depends on the ASS algorithm but also on the audio source mixture scenarios. Therefore, this research addresses a single channel audio source separation problem where the proposed auditory scene is an overlapping multi-lingual multi-dialect speech mixture, which is an inherent characteristic of a cocktail party scenario in India. According to the constitution and census, India is a country with a population of 1.3 billion, speaking 22 official languages and 19,500 languages.

The novel contribution of this paper is the separation of audio sources from a mixture scenario comprising different Indian languages using the NMF algorithm. The quality and intelligibility of the separated signals were analysed with respect to signal level metrics and perception level metrics. The languages used are English, Marathi, and Bengali. The databases used for Bengali and Marathi are taken from openSLR (Open Speech and Language Resources) [28], [29], and TSP speech data [30] for English. The evaluation of the signal level metrics for separation performance was carried out by the “Blind Source Separation evaluation (BSS EVAL)” toolkit [31]. Perception level metrics are given by the “Perceptual Evaluation of Speech Quality (PESQ)” [32] and “Short-Time Objective Intelligibility (STOI)” scores [33]. The performance is further assessed and substantiated by human listeners.

The organization of the paper is as follows: section 2 explains the methodology of NMF and performance measures, implementation is elaborated in section 3, followed by section 4 covering results and discussion. It concludes with section 5.

2. Methodology

This section presents an overview of the “audio source separation” algorithm NMF and the measures used to evaluate the separation performance. The evaluation of separation performance is necessary to assess the capability of the algorithm to separate the signal of interest with significant quality and intelligibility. The performance is compared based on the metrics generated by BSS EVAL, PESQ and STOI.

A. *Non-negative Matrix Factorization*

Paatero and Tapper introduced Positive Matrix Factorization in 1994, later named Non-negative matrix factoriza-

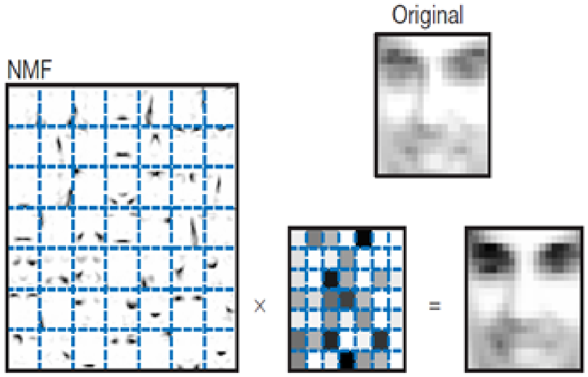


Figure 1. NMF based face reconstruction from image bases and its weight [13]

tion (NMF) [13]. Lee and Seung in 2001 [14], established NMF as a source separation algorithm and demonstrated that though non-negativity constraints limit it, it can learn parts of faces from a facial image database, where a linear combination of these parts or image bases constitutes each face. This is demonstrated in Figure 1, where the parts of faces are the image bases [13].

As earlier mentioned, the competence of NMF was extended to learn semantic features of a text in documents and source separation from mixed signals.NMF separates the target audio source signal (signal of interest) from other interfering speakers or noise or music. It is also capable of separating all the participating signals present in the audio mixture signal. It decomposes the spectrum of mixed audio signal **S** into basically two “non-negative components”. The components are “basis functions matrix” **B** and “weight or activation matrix” **W**.

The approximate factorization **S** ≈ **BW** may be demonstrated elementwise as below [13], [14]:

$$s_{pq} \approx \hat{s}_{pq} = \sum_{k=1}^{L} b_{pk} w_{kq} \tag{1}$$

Where s_{pq} is an element of $\mathbf{S} \in \mathbb{R}_{\geq 0}^{P \times Q}$ factorized into $\mathbf{B} \in \mathbb{R}_{\geq 0}^{P \times L}$ and $\mathbf{W} \in \mathbb{R}_{\geq 0}^{L \times Q}$, all elements of which are subject to the constraints of non-negativity.

$P \in \mathbb{R}_{>0}$ and $Q \in \mathbb{R}_{>0}$ are the axes range of frequencies and time representing the spectrum of the mixed signal **S**, respectively. L $\in \mathbb{R}_{>0}$ is the number of the column basis vectors in **B** $\left(b_{pk}\right.$ is any element of column k) and corresponding weights row wise in **W** (w_{kq} is any element of row k).

The quality of decomposition is quantified using cost functions supported by iterative updates so that it converges to a substantial approximation. The cost function measures the distance or divergence between two non-negative matrices **M** and **N**. **M** and **N**, in this paper, are represented by **S** and **BW**, respectively. The accuracy of the factorization of the object **S** into parts **B** and **W** is, therefore, measured by the cost function convergence. M_{pq} and N_{pq} are elements of **M** and **N**, respectively.

The cost function ”Euclidean distance” (EUC) between **M** and **N** [14] is given by,

$$||\mathbf{M} - \mathbf{N}||^2 = \sum_{pq} \left(M_{pq} - N_{pq}\right)^2 \tag{2}$$

Some cost functions are given the name divergence between **M** and **N**,

$$\mathrm{div}(\mathbf{M}||\mathbf{N}) = \sum_{pq} \left(M_{pq} \log \frac{M_{pq}}{N_{pq}} - M_{pq} + N_{pq}\right) \tag{3}$$

”Equation 3 represents Kullback-Leibler divergence” (KL) [14] which approaches relative entropy when $\sum_{pq} M_{pq} = \sum_{pq} N_{pq} = 1$.

The next divergence given below is ”Itakura-Saito” (IS) divergence [34]

$$\mathrm{div}(\mathbf{M}||\mathbf{N}) = \sum_{pq} \left(\frac{M_{pq}}{N_{pq}} - \log \frac{M_{pq}}{N_{pq}} - 1\right) \tag{4}$$

Both the cost functions are nonincreasing, which leads them to converge to a minimum approximate value. The elements of $\mathbf{B}\left(b_{pk}\right)$ and $\mathbf{W}\left(w_{kq}\right)$ are initialized either randomly or using some pre-defined methodology with non-negative values. $\mathbf{B}^T$ and $\mathbf{W}^T$ are the transpose of **B** and **W**, respectively. **1** is a unity matrix. Convergence is achieved by executing the following multiplicative update theorems iteratively:

The Euclidean distance ||**S** − **BW**|| is updated by the following rules [14],

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\mathrm{B}^T}{\mathrm{B}^T} \frac{\mathrm{S}}{\mathrm{BW}} \quad \mathbf{B} \leftarrow \mathbf{B} \circ \frac{\mathrm{SW}^T}{\mathrm{BWW}^T} \tag{5}$$

The divergence div(**S**||**BW**) for KL uses the following rules [14] to update

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\mathbf{B}^T \frac{S}{\mathrm{BW}}}{\mathbf{B}^T \cdot \mathbf{1}} \quad \mathbf{B} \leftarrow \mathbf{B} \circ \frac{\frac{S}{\mathrm{BW}} \mathbf{W}^T}{\mathbf{1} \cdot \mathbf{W}^T} \tag{6}$$

The above expression shows that a factor is multiplied in each iteration step. This factor is set to unity when **S** = **BW**.

The multiplicative update for IS divergence was established by [34].

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\mathbf{B}^T \frac{\mathbf{S}}{(\mathbf{BW})^2}}{\mathbf{B}^T \cdot \frac{\mathbf{1}}{\mathbf{BW}}} \quad \mathbf{B} \leftarrow \mathbf{B} \circ \frac{\frac{\mathbf{S}}{(\mathbf{BW})^2} \mathbf{W}^T}{\frac{\mathbf{1}}{\mathbf{BW}} \cdot \mathbf{W}^T} \tag{7}$$

B. Performance Measures

The performance measures assess the source separation results based on signal and perception level metrics. Signal level metrics represent the separation performance quality quantitatively by establishing distortion measures between the true and estimated (or separated) sources. However, a human listener is the best judge to assess the separation performance. Therefore, much research has been made to quantify the listener's perception towards a speech signal (in this case, a separated speech signal), leading to the perception level metrics. The perception level metrics are further divided into two categories: (1) objective metrics like PESQ and STOI, which were developed for quality and intelligibility, respectively, (2) subjective metrics, which are obtained by conducting listening tests.

Signal level metrics were evaluated by the BSS Eval toolkit to quantify the amount of speech enhancement or interference mitigation. According to [31], the separated or estimated source $\hat{\mathbf{s}}$ is expressed as a sum of the target source $\mathbf{s}_{target}$ and three types of error as follows:

$$\hat{\mathbf{s}} = \mathbf{s}_{target} + \mathbf{e}_{interf} + \mathbf{e}_{noise} + \mathbf{e}_{artif} \tag{8}$$

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

Energy 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:

$$SDR := 10\log_{10}\frac{\left\|\mathbf{s}_{target}\right\|^2}{\left\|\mathbf{e}_{interf} + \mathbf{e}_{noise} + \mathbf{e}_{artif}\right\|^2} \tag{9}$$

$$SIR := 10\log_{10}\frac{\left\|\mathbf{s}_{target}\right\|^2}{\left\|\mathbf{e}_{interf}\right\|^2} \tag{10}$$

$$SAR := 10\log_{10}\frac{\left\|\mathrm{s}_{target} + \mathrm{e}_{interf} + \mathrm{e}_{noise}\right\|^2}{\left\|\mathrm{e}_{artif}\right\|^2} \tag{11}$$

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 signal. The SAR measures the unwanted energy present in the signal that is not part of either the target or interfering audio signals. The SDR combines the SIR and SAR into one measurement [35].

The human speech intelligibility score is said to be represented by the STOI score. The value range of STOI is typically between 0 and 1. It is computed by obtaining the correlation of short-time temporal envelopes between the clean and separated speech.

The PESQ score gives objective speech quality, a standard metric recommended by the International Telecommunication Union (ITU). The loudness spectrum of a clean reference signal and a separated signal produced by applying auditory transform by PESQ are compared to produce a score in a range of -0.5 to 4.5, which corresponds to the perceptual mean opinion score (MOS) prediction.

3. Implementation

The basic model used for "audio source separation" is shown in Figure 2 Speech signals of different languages were mixed and estimated separately using supervised NMF. The procedure is defined by a training and testing phase. The separated signals are also addressed as estimated signals in this paper. The dataset, steps of experimentation, and the evaluation methods of the same are given below:

A. Dataset

This investigation was evaluated on the synthetic mixtures of two speech signal sources, mainly taken from English, Bengali, and Marathi speech databases. TSP speech database was considered for English language consisting of over 1400 utterances spoken by 24 speakers (12 males, 12 females) developed by Peter Kabal [30], Department of Electrical and Computer Engineering McGill University, Montreal. Bengali and Marathi multi-speaker speech databases are taken from openSLR (Open Speech and Language Resources) [28], [29] developed by Google containing 1366 and 1569 utterances by 9 male and 9 female speakers, respectively. The dataset chosen for the training and testing phase for all the languages are listed below:

1) For English-English speech mixtures, 12 speakers (6 females and 6 males) utterances were selected. Out of these, 10 utterances (5 females and 5 males) are chosen for the training phase, and 1 speaker speech signal, each from both female and male, forms the test data. The utterance of 299 sentences spoken by these 5 female speakers was appended to generate 11 minutes 34 seconds of the clean female speech signal. Similarly, 301 utterances spoken by 5 male speakers were appended to generate 11 minutes 44 seconds of the clean male speech signal. During the testing phase, different speakers and sentences were chosen to make it speaker-independent. The target speech signal is the concatenation of speech signals by 1 female speaker with 60 utterances of 2 minutes 28 seconds. The masker speech signal is the concatenation of 1 male speaker with 60 utterances of 2 minutes 10 seconds.
2) Similarly, for English-Bengali speech mixtures, 6 Bengali speakers (male) are chosen for the training

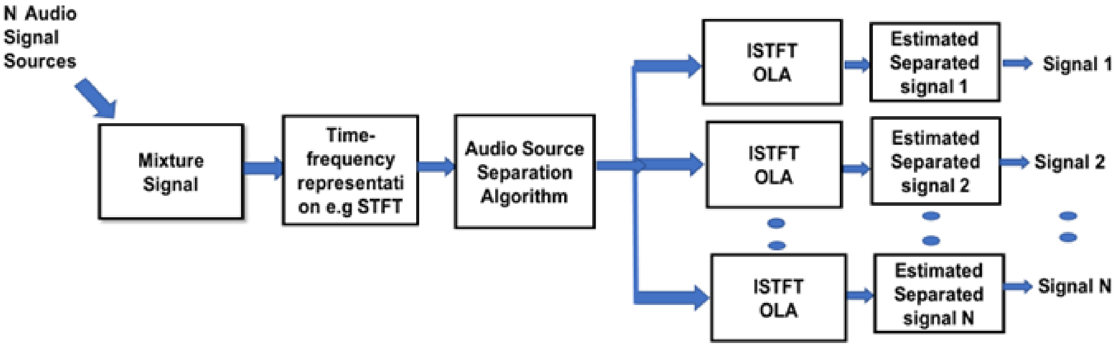

N Audio Signal Sources
Mixture Signal
Time-frequency representation e.g STFT
Audio Source Separation Algorithm
ISTFT OLA
ISTFT OLA
ISTFT OLA
Estimated Separated signal 1
Estimated Separated signal 2
Estimated Separated signal N
Signal 1
Signal 2
Signal N

Figure 2. Audio Source Separation Model

phase with 90 utterances and 1 Bengali speaker (male) with 15 utterances for the testing phase. The English dataset for speakers (female) remains the same as mentioned in point 1.

3) 6 Marathi speakers (female) are chosen for the training phase with 60 utterances and 1 Marathi speaker (female) with 15 utterances during the testing phase for English-Marathi speech mixtures. The English dataset for speakers (female) remains the same as mentioned in point 1.
4) For Bengali-Marathi speech mixtures, 6 Bengali speakers (male) are chosen for the training phase with 90 utterances and 1 Bengali speaker (male) with 15 utterances for the testing phase. 6 Marathi speakers (female) are chosen for the training phase with 90 utterances and 1 Marathi speaker (female) with 14 utterances for the testing phase.

All the audio signals categorized for the training and testing phase were sampled at 16KHz. The test data were augmented by digitally adding target speech signal to masker speech signal with target-to-masker ratios (TMRs) of -10, -5, 0, and 5 and 10 dB. 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 for both speech sources was fixed at 50.

B. *Experimental steps*

As proposed in Figure 2, the ASS algorithm used supervised NMF on the time-frequency representation of the proposed multi-lingual multi-speaker speech scenario. The cost functions with their multiplicative updates used are EUC [14], KL [14], and IS [34]. The separation of different language speech signals, i.e., the target and the masker signals, is obtained by executing the audio source separation algorithm on the synthetic mixtures from the following two speech signal sources:

1) Same English sentences uttered by speakers. (female and male), respectively.
2) Different English sentences uttered by speakers. (female and male), respectively.
3) English and Bengali sentences uttered by female and male speakers, respectively.
4) English and Marathi sentences uttered by male and female speakers, respectively.
5) Marathi and Bengali sentences uttered by female and male speakers, respectively.

The NMF algorithm is given below, and the Python program is available in [36]:

Algorithm 1: NMF

```
Input: non-negative matrix S
Output: non-negative matrices $\mathbf{B}$ and $\mathbf{W}$ such that $\mathbf{S} \approx \mathbf{BW}$
Initialize $\mathbf{B}$ and $\mathbf{W}$ randomly with non-negative values
i=0
Compute cost function EUC or KL or IS
While cost function does not converge do
    $\mathbf{B}^{\mathrm{i}}$ =Update $\mathbf{B}\left(\mathbf{S}, \mathbf{B}^{(\mathrm{i}-1)}, \mathrm{W}^{(\mathrm{i}-1)}\right)$
    $\mathbf{W}^{\mathrm{i}}$ = Update $\mathbf{W}\left(\mathbf{S}, \mathbf{B}^{\mathrm{i}}, \mathbf{W}^{(\mathrm{i}-1)}\right)$
    $\mathrm{i} = \mathrm{i} + 1$
End While
```

In the above algorithm, the cost functions EUC represents Euclidean distance, KL represents Kullback-Leibler divergence and IS represents Itakura Saito divergence. The pseudo-code for the workflow shown in Figure 2 is given below:

Input: Mixed-signal S in time-frequency representation. Output: Separated sources Signal 1 (s1), Signal 2 (s2).

Step 1: Perform NMF on individual speech sources (different languages) to obtain their “basis functions” in the training phase.

Step 2: The “basis functions matrix” obtained after factorization during the training phase is passed onto the testing stage.

Step 3: Only the “weight or activation matrix” is estimated while factorizing the mixed-signal S during the testing phase.

Step 4: A specific source is separated by multiplying the “basis functions” and its corresponding “weights”.

Step 5: Inverse Fourier transform and “overlap-and-add” (OLA) routine are used to obtain the time-domain signal of a separated source s1 and s2 from the corresponding estimated magnitude spectrum and the mixed-signal phase.

Apart from the above, the experiments were tried with the same gender. For the English language, both male-male and female-female, female-female for Marathi and male-male for Bengali combinations were investigated. Python programming language was used for the NMF algorithm with multiplicative updates. Parselmouth, PRAAT in Python [37] was used for the spectrograms.

C. Evaluation

The source separation results were evaluated using the signal level metrics using BSS_EVAL tool which measures the performance of separation using the parameters, source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifact ratio (SAR). Objective metrics like PESQ and STOI measures the quality and intelligibility of the separated speech, respectively, and subjective metrics are obtained by conducting listening tests.

Subjective listening tests were performed by requesting 10 human listeners. The speech signals were played on a loudspeaker in a typical environmental condition. All the listeners knew the English language and Marathi language. 6 listeners knew all three languages. The age of the listeners (both female and male) ranges from 22 yrs. to 54 yrs. The subjective tests included the mean opinion score for speech intelligibility (MOS-I) and speech quality (MOS-Q). For MOS-I, the human listeners were asked to rate the accuracy to hear what is being said. For MOS-Q, they were asked to rate the goodness of the speech quality. Therefore, the human listeners rate the estimated speech signal into one of the five quality categories, both for quality and intelligence, which is mapped to Absolute Category Rating (ACR, P.800) [38], shown in Table I.

TABLE I. SUBJECTIVE LISTENING TEST METRICS: SPEECH QUALITY AND INTELLIGIBILITY.

Rating	Speech Quality Category (P.800)	Speech Quality MOS-Q	Speech Intelligibility MOS-I
5	Excellent	Very good	Perceptible
4	Good	Good	Perceptible with slight difficulty
3	Fair	Normal	Perceptible when no overlap
2	Poor	Bad	Few words clear
1	Bad	Very Bad	Not clear

Python was used for all the performance metrics except PESQ. MATLAB was used for PESQ results.

4. Results and Discussion

This section is divided into three subsections: speech separation, signal level metrics, and perception level metrics. The speech separation section shows the separation procedure with the help of spectrograms, as shown in Figure 3. Signal level metrics compare BSS_EVAL values for the separated signals from the multi-lingual speech mixture. Perception level metrics compare the objective metrics using PESQ and STOI and subjective metrics using human listeners.

Though speech combinations of speakers male-male and female-female for English language, speakers female-female for Marathi and speakers male-male for Bengali were investigated, the separation performance results of these combinations were poor. Therefore, the results discussed here comprises only female and male speech combinations in a mixed signal.

A. Speech Separation

The most appropriate way to understand the separation of the target and the masker speeches from a mixed speech using the NMF algorithm is by means of spectrograms. A mixed speech signal consisting of Marathi and Bengali speaker sentences (female and male) with TMR 0 dB is chosen. The signal strength, which varies over time at different frequencies, is shown in the spectrogram by the dB bar added to it. The spectrograms are truncated to 4 seconds to represent the speech signals represented in Figure 3.

The figure depicts the mixed spectrogram mixed at 0 dB, the original target, and the masker. It also shows the separated target and the masker, respectively, at 0 dB. The fundamental frequency is also added to the spectrograms to get a clear idea about the interference speech signal’s traces in the separated target speech.

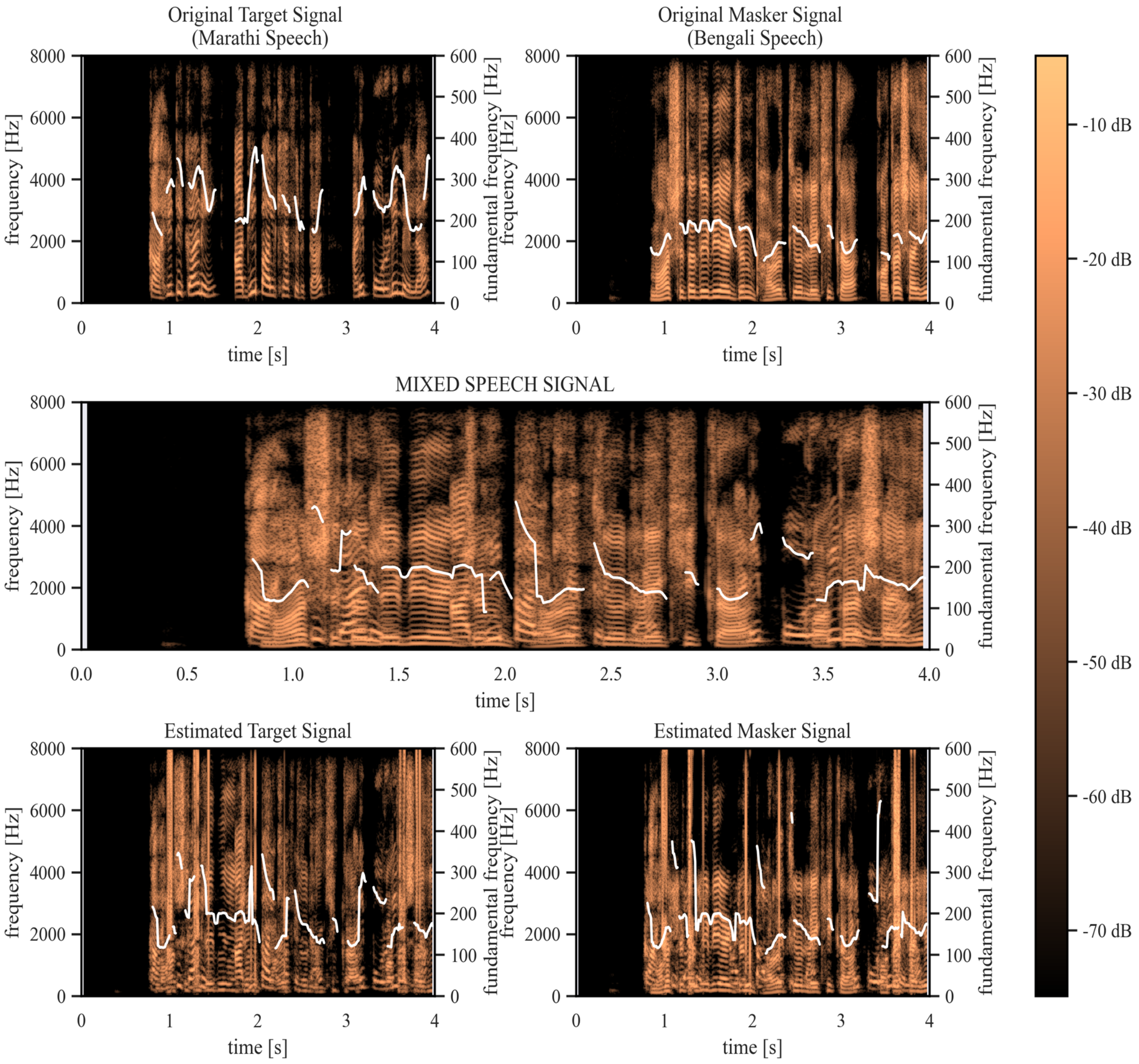


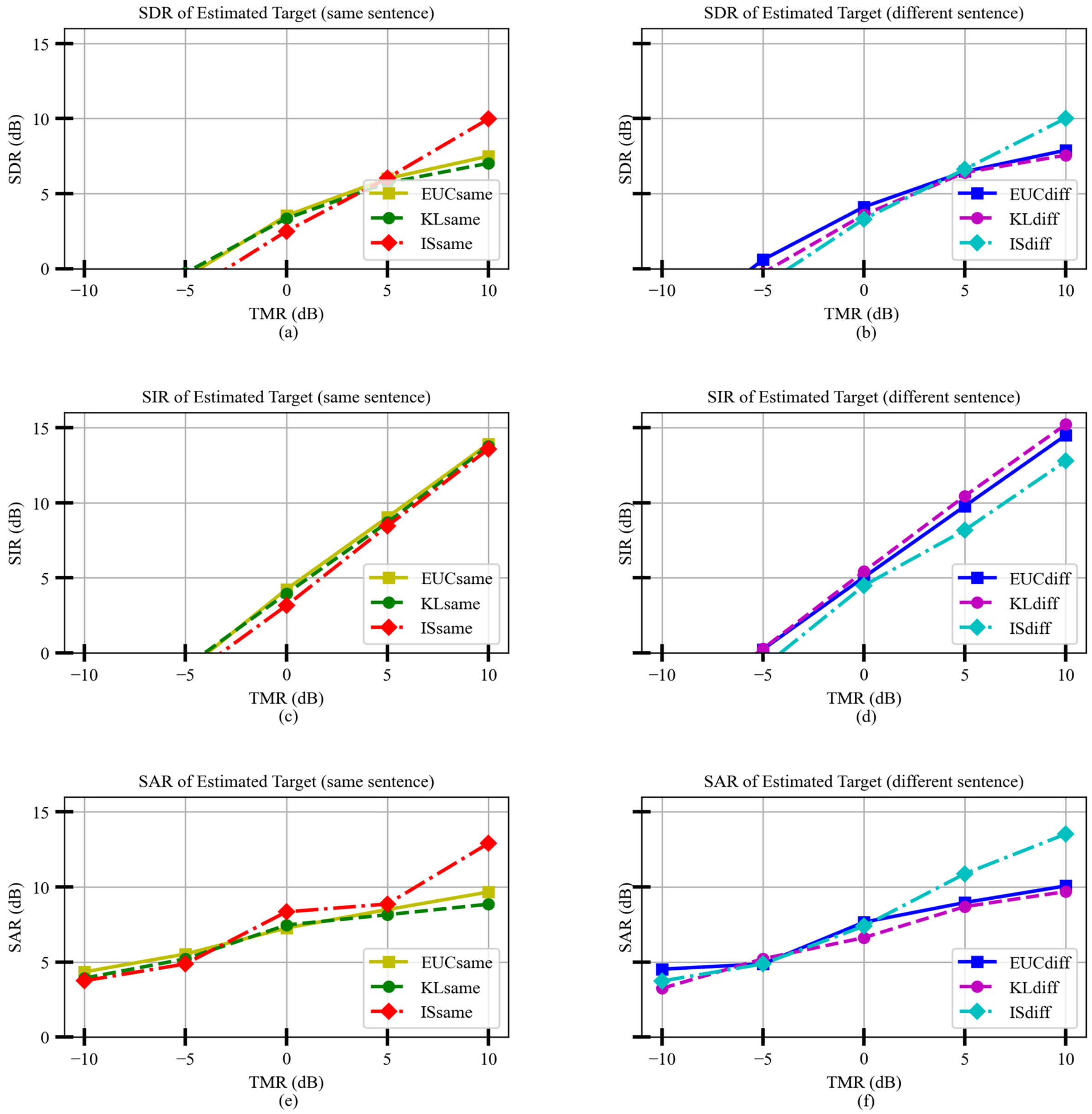
Figure 3. Separation of the target and the masker speeches from a speech mixture signal mixed at 0 dB TMR. Original source speech (target and masker), mixed speech, the estimated speech (target and masker) after separation

It is observed from the spectrograms that the separated Marathi speech (target speaker) spoken by a female speaker is having residual traces of Bengali speech (masker speaker) spoken by a male speaker. Similarly, the separated masker is having very few traces of the target. This result is validated by the human listeners in the sections of perception level metrics.

B. *Signal level metrics*

This section compares BSS EVAL values for the estimated target speech signal after separation. It is done in two ways: (1) the metrics for the language combinations are tabulated in Table II, (2) the metrics for the mixed speech signal consisting of same and different English sentences uttered by a female and male speaker is plotted against the different TMRs. The different TMRs highlighted are -10, -5, 0, 5 and 10 dB, respectively. NMF performance comparisons using different cost functions are shown in the plots. This comparison identifies the cost function giving the best results for the same and different sentence speech mixtures. The SDR, SIR and SAR values of the estimated or separated target audio source for NMF-EUC, NMF-KL and NMF-IS on an overlapped speech signal comprising same and different sentence utterances by English female and male speakers, are given in Figure 4, respectively.

It is observed that in both the SDR values, KL divergence gives better results in lower dB TMR, and IS divergence gives good results in higher dB TMRs. At 0 dB, Euclidean distance is giving the highest SDR. It is also observed that the SDR results of an estimated target from



(a)

(b)

(c)

(d)

(e)

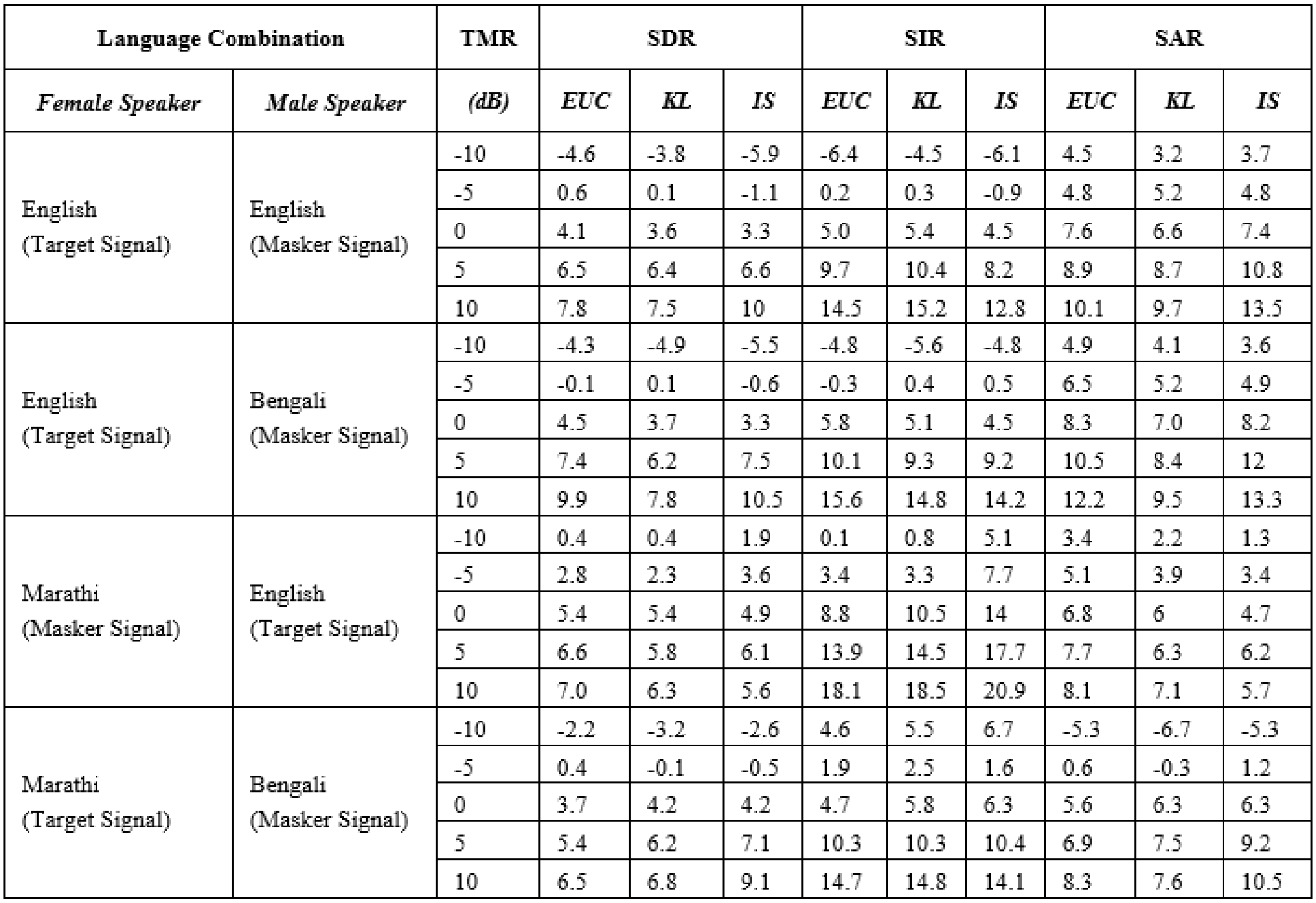
(f)

Figure 4. Performance comparison of NMF-EUC, NMF-KL and NMF-IS on an overlapped speech signal comprising same and different sentence utterances by English speakers (female and male), (a) SDR (same sentence), (b) SDR (Different sentence), (c) SIR (same sentence), (d) SIR (Different sentence), (e) SAR (same sentence), (f) SAR (Different sentence). The y-axis scale for the figures (b), (c), (d) , (e) and (f) are same as in figure (a)

different sentence mixed-signal are more (nearly 1 dB) than the same sentence mixed-signal. The results of SIR are best shown by KL divergence in both cases, and it is nearly 2 to 3 dB more for different sentence mixed signal. SAR results in both cases show that the IS divergence depicts the best results. For different sentence mixed-signal, the SAR for the estimated target is 2 dB more than the same sentence mixed-signal at 5 dB TMR.

Therefore, in all the cases, the estimated target's separation performance is better when the mixed English speech signal consists of different sentences uttered by females and males, respectively. To put it another way, it is easier to separate from different sentence mixed signals than the same sentence mixed signals. The separation, therefore, improves when the sources are less related. A study based on vowels [2] was conducted on a similar line, which showed that the results of separating the vowels from a synthetic audio mixture of two dissimilar vowels were better than two similar vowels uttered by a female-male

TABLE II. PERFORMANCE COMPARISON (BSS) BETWEEN NMF DIVERGENCES OF THE ESTIMATED TARGET FOR ALL COMBINATIONS OF MIXED SPEECH SIGNAL WITH DIFFERENT TMRS FOR TWO SPEAKERS



Language Combination		**TMR**	**SDR**			**SIR**			**SAR**		
Female Speaker	***Male Speaker***	***(dB)***	***EUC***	***KL***	***IS***	***EUC***	***KL***	***IS***	***EUC***	***KL***	***IS***
English (Target Signal)	English (Masker Signal)	-10	-4.6	-3.8	-5.9	-6.4	-4.5	-6.1	4.5	3.2	3.7
		-5	0.6	0.1	-1.1	0.2	0.3	-0.9	4.8	5.2	4.8
		0	4.1	3.6	3.3	5.0	5.4	4.5	7.6	6.6	7.4
		5	6.5	6.4	6.6	9.7	10.4	8.2	8.9	8.7	10.8
		10	7.8	7.5	10	14.5	15.2	12.8	10.1	9.7	13.5
English (Target Signal)	Bengali (Masker Signal)	-10	-4.3	-4.9	-5.5	-4.8	-5.6	-4.8	4.9	4.1	3.6
		-5	-0.1	0.1	-0.6	-0.3	0.4	0.5	6.5	5.2	4.9
		0	4.5	3.7	3.3	5.8	5.1	4.5	8.3	7.0	8.2
		5	7.4	6.2	7.5	10.1	9.3	9.2	10.5	8.4	12
		10	9.9	7.8	10.5	15.6	14.8	14.2	12.2	9.5	13.3
Marathi (Masker Signal)	English (Target Signal)	-10	0.4	0.4	1.9	0.1	0.8	5.1	3.4	2.2	1.3
		-5	2.8	2.3	3.6	3.4	3.3	7.7	5.1	3.9	3.4
		0	5.4	5.4	4.9	8.8	10.5	14	6.8	6	4.7
		5	6.6	5.8	6.1	13.9	14.5	17.7	7.7	6.3	6.2
		10	7.0	6.3	5.6	18.1	18.5	20.9	8.1	7.1	5.7
Marathi (Target Signal)	Bengali (Masker Signal)	-10	-2.2	-3.2	-2.6	4.6	5.5	6.7	-5.3	-6.7	-5.3
		-5	0.4	-0.1	-0.5	1.9	2.5	1.6	0.6	-0.3	1.2
		0	3.7	4.2	4.2	4.7	5.8	6.3	5.6	6.3	6.3
		5	5.4	6.2	7.1	10.3	10.3	10.4	6.9	7.5	9.2
		10	6.5	6.8	9.1	14.7	14.8	14.1	8.3	7.6	10.5

combination. As earlier mentioned, a comparison of signal level metrics for all the combinations for the estimated target is tabulated in Table II.

The BSS metrics, i.e., SDR, SIR, and SAR values, increase with increasing TMR for the estimated target. Similarly, the values decrease with decreasing TMR. The separation is, therefore, difficult for the weaker signal. Comparing the values at 0 dB shows that Euclidean distance has faired better for all the combinations than the other divergences except few instances. Another observation is that one of the signals is estimated better than the other at 0 dB.

As the separation performance of NMF using EUC, KL, and IS divergences for overlapped speech mixtures of different Indian languages is not available in the literature, it is compared with the English language mixed speech signal, the BSS results of which is verified with similar research [4]. The current investigation of NMF on speech mixtures is also compared with previous results as in Figure 5 [25], [4].

At 0 dB TMR, the SDR of separated target source from English-Bengali and English-Marathi speech mixture is 0.4 and 1.3 dB higher than English-English speech mixed signals, respectively. Similarly, SIR values for English-Bengali and English-Marathi mixed signals are 0.8 and 3.8 dB, respectively, higher than English-English mixed speech. Though the SAR value for English-Bengali mixed speech is higher by 0.7, it is 0.7 lower in English-Marathi mixed speech than English-English mixed speech implying that more artifacts are produced in the case of English-Marathi speech mixture separation. Audible artifacts may be introduced in the reconstructed time-domain signal due to the mixed audio signal phase. At 0 dB, all the performance measures for the estimated speech signal for Marathi-Bengali mixed-signal reduce subsequently.

To summarize, the higher SIR and SDR values highlight that the estimated speeches after separation are less interfered with and distorted by the masker speaker for English-Bengali and English-Marathi combination than English-English and Marathi-Bengali mixed speech signal.

The above result may be explained by the fact that all

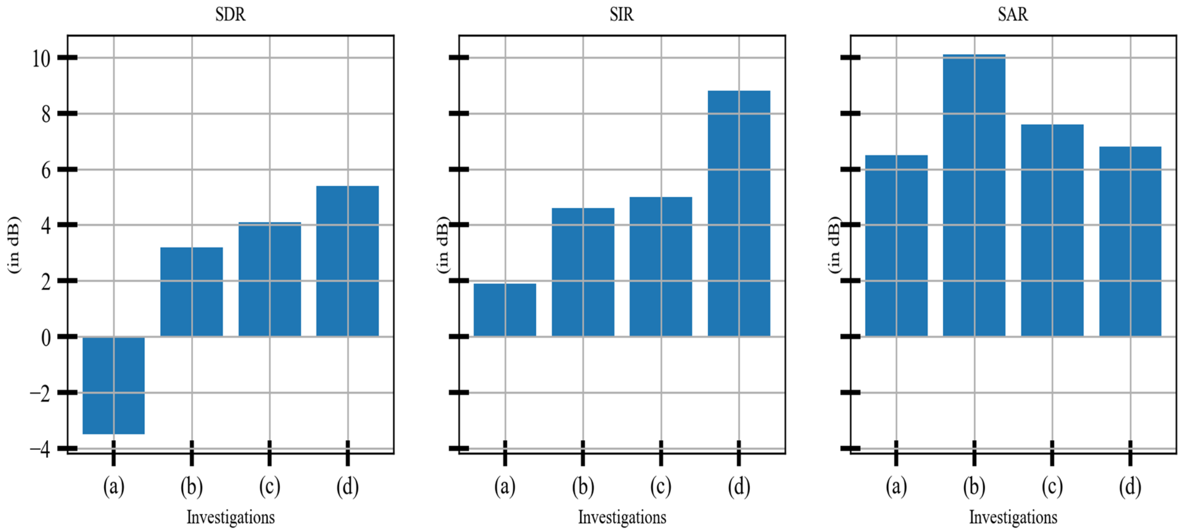


Figure 5. SDR, SIR, and SAR comparison of (a) NMF-EUC(sparse) [25], (b) NMF-EUC [4], (c) NMF-EUC, current investigation on an overlapped speech signal comprising same language speeches by English female and male speakers and (d) NMF-EUC, current investigation on an overlapped speech signal comprising different language by English male and Marathi female speakers. The y-axis scale for all the figures is same

Indo-Aryan languages like Marathi and Bengali have more aspirated consonants than English, which are produced with an audible expulsion of breath, whereas the unaspirated are pronounced with minimal breath. Another contrast is between dental and retroflex consonants; the upper front teeth are touched by the tongue in dental consonants, whereas the tip of the tongue is curled upwards against the palate in case of retroflex consonants. Indian retroflex sound is produced when the tongue is released from this position [39], [40]. Voice onset time (VOT), which is known to vary with place of articulation, has been used to differentiate stop categories across languages [39]. Therefore, the sources are less related in the case of English-Bengali and English-Marathi mixed signals than the other two language combinations, may explain the improvement in separation performance. The results are all validated by following perception level metrics.

C. Perception level metrics

This section is divided into two sections: objective metrics using PESQ and STOI and subjective metrics using human listeners.

PESQ and STOI scores for the estimated target and masker are tabulated in Table III. The scores are calculated only for the estimated target and masker separated by NMF with Euclidean distance. For lower TMRs, the masker is estimated better, and for higher TMRs, the target is estimated better. Comparing the 0 dB TMR for all the combinations shows that the estimated target has better PESQ and STOI values for English-Bengali and English-Marathi than the other two, which is in line with the BSS values in the previous section.

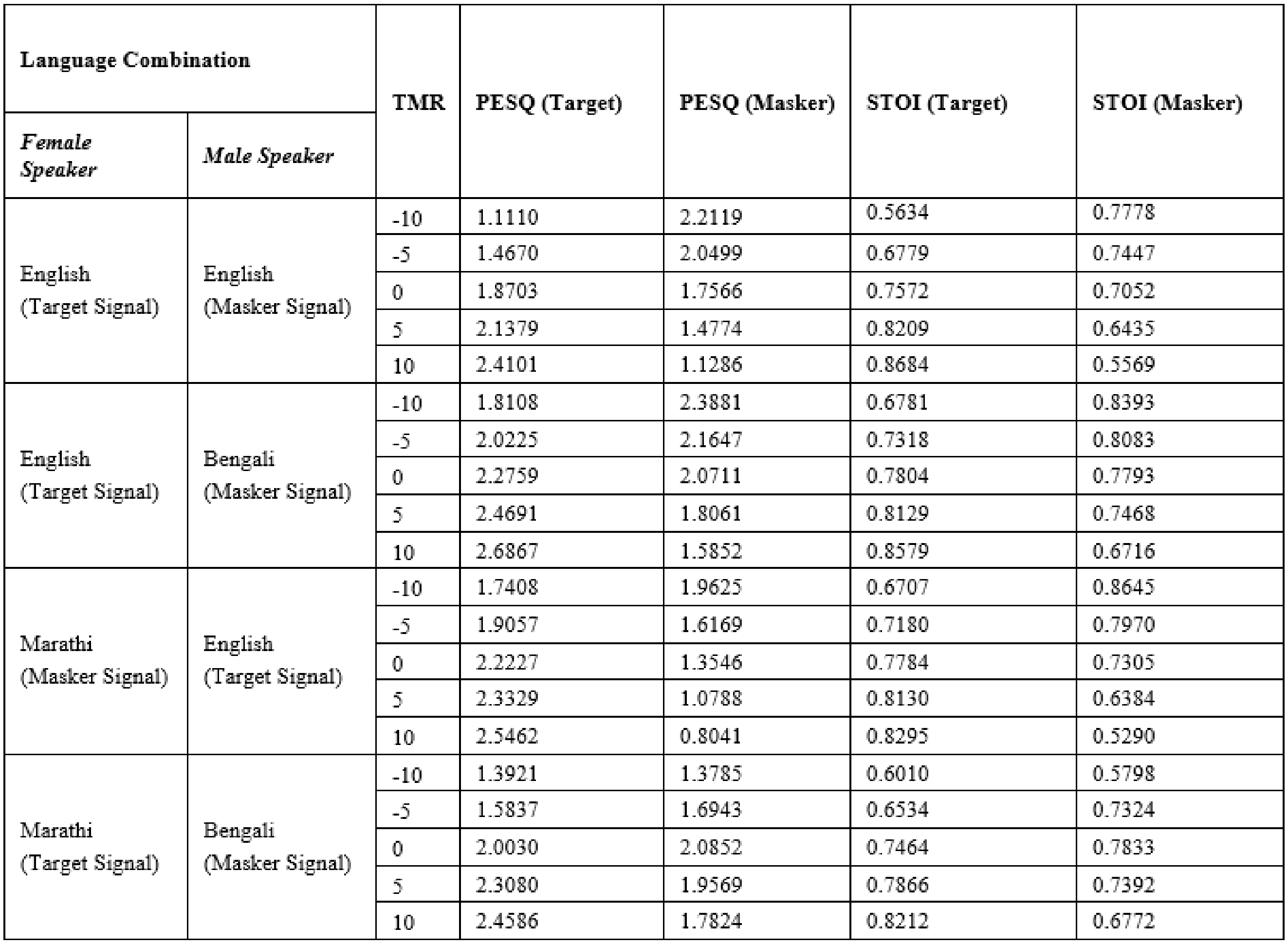
The quality and intelligibility of the target signal, i.e., PESQ (target) and STOI (target), for higher TMRs such as 5 dB or 10 dB show results in the range of 2.3 to 2.5 and 0.78 to 0.87, respectively, which is significantly better. In accordance, PESQ (masker) and STOI (masker) values for masker signals are much lesser for higher TMRs. Similarly the results for lower TMRs are better for masker signals.

Subjective human listening tests conducted are tabulated in Table IV. As perceived by human listeners, MOS-Q and MOS-I identify with quality and masker signal in the estimated target after separation varies with speech combinations of different language in the mixture signal, individuals are asked to give their opinion as per Table I on a scale of 1(low) to 5(high) whether the sentences are of good quality and understood well. The higher the scores, the better is the separation performance leading to a reasonable estimation of the signal. The average value of all the listeners' opinion is taken to get the values of MOS-Q and MOS-I for each of the speech under the test. The values for MOS-Q and MOS-I for target signal for all the combinations are in the range of 4 to 5 and 3 to 5, respectively which suggests normal to very good quality.

According to human listeners, the speeches separated from English-Marathi mixed signal combination depicts better quality than the other language combinations. The subjective listening test scores validate the objective metrics PESQ and STOI.

In the experiment involving mixed-signal speeches of the same sentence and different sentence uttered by male and female English speakers, the results demonstrated that speeches were better separated in case of different sentence combinations. Similarly, the speeches separated from

TABLE III. Performance comparison (PESQ and STOI) of the (NMF EUC) estimated target and masker for all combinations of mixed speech signal with different TMRs



Language Combination		**TMR**	**PESQ (Target)**	**PESQ (Masker)**	**STOI (Target)**	**STOI (Masker)**
Female Speaker	***Male Speaker***					
English (Target Signal)	English (Masker Signal)	-10	1.1110	2.2119	0.5634	0.7778
		-5	1.4670	2.0499	0.6779	0.7447
		0	1.8703	1.7566	0.7572	0.7052
		5	2.1379	1.4774	0.8209	0.6435
		10	2.4101	1.1286	0.8684	0.5569
English (Target Signal)	Bengali (Masker Signal)	-10	1.8108	2.3881	0.6781	0.8393
		-5	2.0225	2.1647	0.7318	0.8083
		0	2.2759	2.0711	0.7804	0.7793
		5	2.4691	1.8061	0.8129	0.7468
		10	2.6867	1.5852	0.8579	0.6716
Marathi (Masker Signal)	English (Target Signal)	-10	1.7408	1.9625	0.6707	0.8645
		-5	1.9057	1.6169	0.7180	0.7970
		0	2.2227	1.3546	0.7784	0.7305
		5	2.3329	1.0788	0.8130	0.6384
		10	2.5462	0.8041	0.8295	0.5290
Marathi (Target Signal)	Bengali (Masker Signal)	-10	1.3921	1.3785	0.6010	0.5798
		-5	1.5837	1.6943	0.6534	0.7324
		0	2.0030	2.0852	0.7464	0.7833
		5	2.3080	1.9569	0.7866	0.7392
		10	2.4586	1.7824	0.8212	0.6772

English-Bengali, and English-Marathi mixed speech signal combination shows better results than the combination of Marathi-Bengali. It is observed therefore, that the speech is separated better when the mixed signal consists of English and an Indian language.

The sources are less related in English-Bengali and English-Marathi mixed signals than the other two combinations, as explained earlier. Therefore, the experiments' results highlight that applying the ASS algorithm on a mixed-signal containing audio sources of different language combination enhances the performance of separation. The combinations using the Marathi language show less SAR values, which need more in-depth investigation.

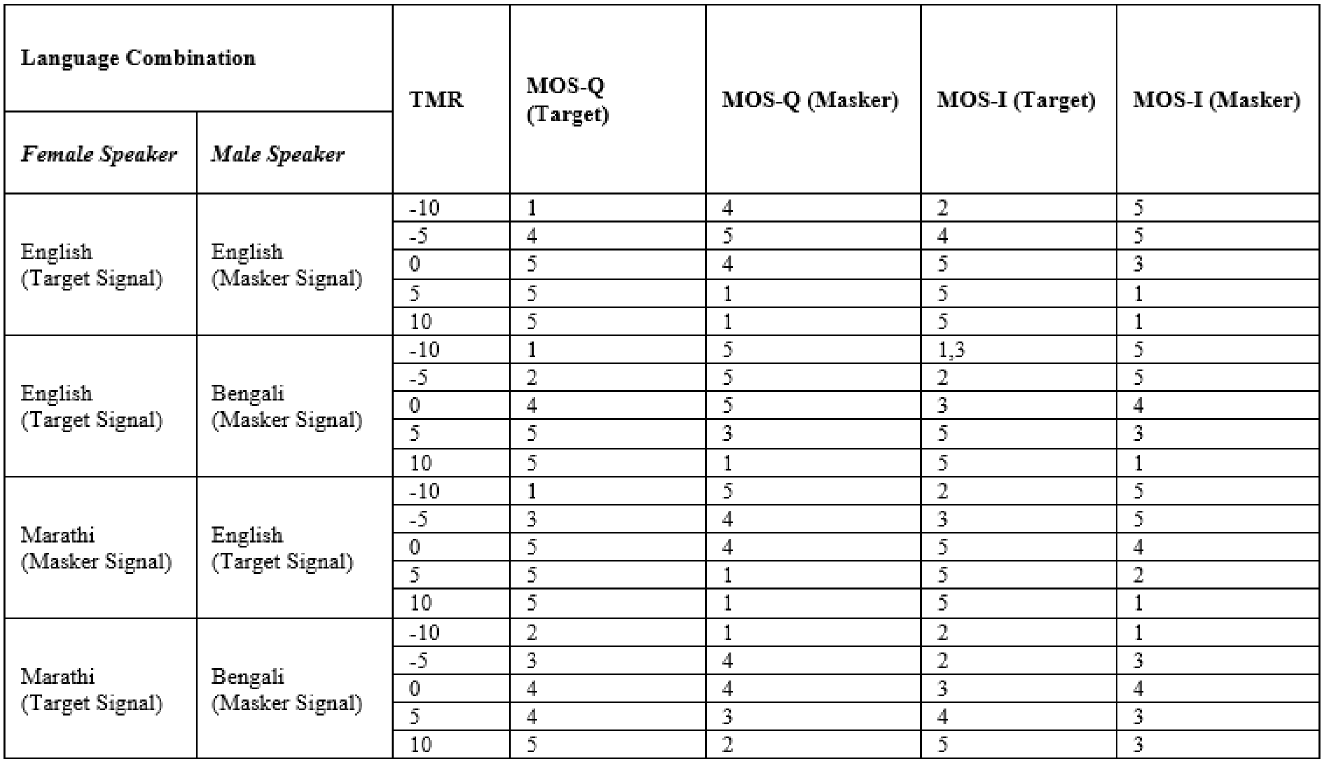
NMF on a multi-lingual mixed speech signal as a preprocessor may be added to a speech recognition module. This will mitigate the interference caused by overlapping speeches comprising languages other than the target language.

5. Conclusion

Any cocktail party scenario in India would comprise several people speaking different languages. Therefore, this paper reports investigation on NMF based source separation for a multi-lingual overlapped speech signal scenario. The signal and perception level metrics are observed and analyzed, which includes perceptual data from human listeners also. The results at 0 dB TMR shows that the SDR of separated target source from English-Bengali and English-Marathi speech mixture is 0.4 and 1.3 dB higher than English-English speech mixed signals, respectively. Similarly, SIR values for English-Bengali and English-Marathi mixed signals are 0.8 and 3.8 dB, respectively, higher than English-English mixed speech. Therefore, an improvement in separating sources from mixed speech signals with different language combinations than the same language is the highlight of this investigation.

Several companies like Google and Microsoft are coming up with Indian language speech databases which are not

TABLE IV. HUMAN LISTENERS – MOS-Q AND MOS-I



Language Combination		**TMR**	**MOS-Q (Target)**	**MOS-Q (Masker)**	**MOS-I (Target)**	**MOS-I (Masker)**
Female Speaker	***Male Speaker***					
English (Target Signal)	English (Masker Signal)	-10	1	4	2	5
		-5	4	5	4	5
		0	5	4	5	3
		5	5	1	5	1
		10	5	1	5	1
English (Target Signal)	Bengali (Masker Signal)	-10	1	5	1,3	5
		-5	2	5	2	5
		0	4	5	3	4
		5	5	3	5	3
		10	5	1	5	1
Marathi (Masker Signal)	English (Target Signal)	-10	1	5	2	5
		-5	3	4	3	5
		0	5	4	5	4
		5	5	1	5	2
		10	5	1	5	1
Marathi (Target Signal)	Bengali (Masker Signal)	-10	2	1	2	1
		-5	3	4	2	3
		0	4	4	3	4
		5	4	3	4	3
		10	5	2	5	3

in abundance. Nowadays, the speech recognition module is modified to adapt to various Indian languages, and thus separation of audio signals comprising different languages may add up to its benefit. This experimentation was conducted to explore an Indian scenario and may be further explored by newer models of NMF, and the separation performance may be investigated using recognition tools.

References

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, Sep. 1953, publisher: Acoustical Society of America. https://asa.scitation.org/doi/10.1121/1.1907229

[2] N. C. Nag and M. Shah, "Understanding Basis Functions for Vowels Based on Non-negative Matrix Factorization," in *2017 International Conference on Nascent Technologies in Engineering (ICNTE)*, Jan. 2017, pp. 1–6.

[3] A. Abad Gareta, "A Multi-microphone Approach to Speech Processing in a Smart-room Environment," Ph.D. Thesis, Universitat Politècnica de Catalunya, Jun. 2007. http://www.tdx.cat/handle/10803/6906

[4] G. Bao, Y. Xu, and Z. Ye, "Learning a Discriminative Dictionary for Single-Channel Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1130–1138, Jul. 2014.

[5] S. Sajid, A. Javed, and A. Irtaza, "An Effective Framework for Speech and Music Segregation," *IAJIT*, vol. 17, no. 4, pp. 507–514, Jul. 2020. https://iajit.org/PDF/July%202020,%20No.%204/16971.pdf

[6] F. Li and M. Akagi, "Combining F0 and non-negative constraint robust principal component analysis for singing voice separation," *Signal Process.*, 2020.

[7] K. H. Thanoon, S. Q. Hasan, and O. I. Alsaif, "Biometric Information Based on Distribution of Arabic Letters According to Their Outlet." *International Journal of Computing and Digital Systems*, vol. 9, no. 5, pp. 981–992, Sep. 2020, publisher: University of Bahrain, Deanship of Graduate Studies and Scientific Research. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=2210142X&v=2.1&it=r&id=GALE%7CA636969267&sid=googleScholar&linkaccess=abs

[8] A. Jasim, J. Mahmood, and R. Al-Waily, "Design and Implementation of a Musical Water Fountain Based on Sound Harmonics Using IIR Filters," *International Journal of Computing and Digital Systems*, vol. 9, Mar. 2020.

[9] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[10] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994, publisher: Elsevier.

[11] T.-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind Source Separation of More Sources Than Mixtures Using Overcomplete Representations," *Signal Processing Letters, IEEE*, vol. 6, pp. 87–90, Apr. 1999.

[12] J. Wang, S. Guan, S. Liu, and X.-L. Zhang, “Minimum-volume Multichannel Nonnegative Matrix Factorization for Blind Source Separation,” *arXiv:2101.06398 [cs, eess]*, Mar. 2021. http://arxiv.org/abs/2101.06398

[13] 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, Oct. 1999, bandiera_abtest: a Cg_type: Nature Research Journals Number: 6755 Primary_atype: Research Publisher: Nature Publishing Group. https://www.nature.com/articles/44565

[14] D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization,” in *Advances in Neural Information Processing Systems*, vol. 13. MIT Press, 2001.

[15] P. Mowlaee, M. G. Christensen, and S. H. Holdt Jensen, “Improved single-channel speech separation using sinusoidal modeling,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2010, pp. 21–24.

[16] P. Comon, “Independent component analysis, A new concept?” *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994. https://www.sciencedirect.com/science/article/pii/0165168494900299

[17] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis,” *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020. https://www.frontiersin.org/article/10.3389/frai.2020.00042

[18] G. Phelan, “New perspectives on non-negative matrix factorization for grouped topic models,” Aug. 2020. http://summit.sfu.ca/item/20921

[19] H. Horkous and M. Guerti, “Recognition of Anger and Neutral Emotions in Speech with Different Languages,” *International Journal of Computing and Digital Systems*, vol. 10, pp. 563–574, Apr. 2021, accepted: 2021-03-03T14:33:44Z Publisher: University of Bahrain. https://journal.uob.edu.bh:443/handle/123456789/4145

[20] S. Singh, N. Singh, and D. Chaudhary, “A Survey on Autonomous Techniques for Music Classification based on Human Emotions Recognition,” *International Journal of Computing and Digital Systems*, vol. 9, pp. 433–447, May 2020.

[21] N. N. J. Siphocly, E.-S. M. El-Horbaty, and A.-B. M. Salem, “Top 10 Artificial Intelligence Algorithms in Computer Music Composition.” *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 373–395, Jan. 2021, publisher: University of Bahrain, Deanship of Graduate Studies and Scientific Research. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=2210142X&v=2.1&it=r&id=GALE%7CA653985664&sid=googleScholar&linkaccess=abs

[22] J. Le Roux, J. R. Hershey, and F. Weninger, “Deep NMF for speech separation,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 66–70.

[23] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, “NMF-based speech enhancement incorporating deep neural network,” in *Interspeech 2014*. ISCA, Sep. 2014, pp. 2843–2846.

[24] Z. Chen, *Single Channel auditory source separation with neural network*. Columbia University, 2017.

[25] T. F. Krikke, F. Broz, and D. Lane, “Who Said That? A Comparative Study of Non-Negative Matrix Factorisation and Deep Learning Techniques,” *AAAI 2017 Fall Symposium*, p. 5, 2017.

[26] A. M. S. Ang, N. Gillis, A. Vandaele, and H. D. Sterck, “Non-negative Unimodal Matrix Factorization,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 3270–3274, iSSN: 2379-190X.

[27] M. Charikar and L. Hu, “Approximation Algorithms for Orthogonal Non-negative Matrix Factorization,” in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. PMLR, Mar. 2021, pp. 2728–2736, iSSN: 2640-3498.

[28] F. He, S.-H. C. Chu, O. Kjartansson, C. Rivera, A. Katanova, A. Gutkin, I. Demirsahin, C. Johny, M. Jansche, S. Sarin, and K. Pipatsrisawat, “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*. Marseille, France: European Language Resources Association, May 2020, pp. 6494–6503.

[29] K. Sodimana, K. Pipatsrisawat, L. Ha, M. Jansche, O. Kjartansson, P. D. Silva, and S. Sarin, “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*, 2018, pp. 66–70.

[30] “TSP Lab - Data.” http://www-mmsp.ece.mcgill.ca/Documents/Data/

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

[32] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, May 2001, pp. 749–752 vol.2.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[34] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[35] B. J. King, “_USNew Methods of Complex Matrix Factorization for Single-Channel Source Separation and Analysis,” Thesis, Apr. 2013. https://digital.lib.washington.edu:443/researchworks/handle/1773/22555

[36] “My Drive - Google Drive.” https://drive.google.com/drive/u/1/my-drive

[37] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, Nov. 2018. https://www.sciencedirect.com/science/article/pii/S0095447017301389

[38] “P.800 : Methods for subjective determination of transmission quality.” https://www.itu.int/rec/T-REC-P.800-199608-I

[39] I. Datta, "Morphology of Bengali and Tamil language a contrastive study," *University*, 2008, accepted: 2017-07-10T10:18:41Z Publisher: Kolkata. http://shodhganga.inflibnet.ac.in:8080/jspui/handle/10603/159739

[40] A. Balyan, "Creating an Unlimited Voice Response in Hindi," *University*, 2016. http://shodhganga.inflibnet.ac.in:8080/jspui/handle/10603/183498

Nandini C Nag Nandini C Nag was born in Tamluk, Midnapore, West Bengal, India in 1970. She received the B.E. degree in electronics and communication engineering from the National Institute of Technology, Durgapur, West Bengal, India in 1992 and the M.E. degree in information technology from VESIT, Chembur, University of Mumbai, India in 2006. She is currently pursuing a PhD degree in electronics and telecommunication engineering at Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai, India. Her research interests include speech processing. Prof. Nag is a member of the Indian Society for Technical Education, India.

Milind S Shah Milind S Shah was born in Pen Village, Raigad District, Maharashtra, India, in 1968. He received the B.E. and M.Tech. degrees in electronics engineering from the University of Mumbai and Nagpur University in 1990 and 1993, respectively and the PhD degree in electrical engineering from the Indian Institute of Technology Bombay, Mumbai, India, in 2008. From 1991 to 1995 and 1995 to 1998, he was a Lecturer with the K. E. S. Engineering College and Dr. Babasaheb Ambedkar Technological University, respectively. From 1998 to 2008, he was Assistant Professor with the Fr. C. Rodrigues Institute of Technology. Since 2008, he has been a Professor with the Electronics and Telecommunication Engineering Department, Fr. C. Rodrigues Institute of Technology (affiliated to the University of Mumbai), Navi Mumbai, India. He has published more than 50 articles in various conferences and journals. His research interests include speech and signal processing. Prof. Shah is a senior member of the IEEE, Fellow of the Institution of Engineers, India, and a Fellow of the Institution of Electronics and Telecommunication Engineers, India.