PAN ARCHIVES OF ACOUSTICS
Vol. **42**, No. 2, pp. 287–295 (2017)

DOI: 10.1515/aoa-2017-0031

Frequency Selection Based Separation of Speech Signals with Reduced Computational Time Using Sparse NMF

Yash Vardhan VARSHNEY, Zia Ahmad ABBASI, Musiur Raza ABIDI, Omar FAROOQ

Department of Electronics
Aligarh Muslim University
Aligarh, India
e-mail: {yashvarshneyy, omarfarooq70}@gmail.com, zaabbasi@zhcet.ac.in, abidimr@rediffmail.com

(*received September 2, 2016; accepted December 21, 2016*)

Application of wavelet decomposition is described to speed up the mixed speech signal separation with the help of non-negative matrix factorisation (NMF). It is assumed that the basis vectors of training data of individual speakers had been recorded. In this paper, the spectrogram magnitude of a mixed signal has been factorised with the help of NMF with consideration of sparseness of speech signals. The high frequency components of signal contain very small amount of signal energy. By rejecting the high frequency components, the size of input signal is reduced, which reduces the computational time of matrix factorisation. The signal of lower energy has been separated by using wavelet decomposition. The present work is done for wideband microphone speech signal and standard audio signal from digital video equipment. This shows an improvement in the separation capability using the proposed model as compared with an existing one in terms of correlation between separated and original signals. Obtained signal to distortion ratio (SDR) and signal to interference ratio (SIR) are also larger as compare of the existing model. The proposed model also shows a reduction in computational time, which results in faster operation.

Keywords: sparse NMF; mixed speech recognition; machine learning.

1. Introduction

The problem of mixed signal separation have been attracting researchers for a long time. Non-negative matrix factorisation (NMF) has emerged for the use of source separation (Lee, Seung, 1999; Paatero, Tapper, 1994). Initially used for mathematical computation, NMF has now found its application in the field of various source separations. NMF factorises a two dimensional matrix into its components and their weights. For speech signals, the spectrogram magnitude can be considered as primary two-dimensional matrix. As speech signals are sparse in nature, sparse NMF is applied for factorise it (Benetos *et al.,* 2006; Demir *et al.,* 2013; Schmidt, Olsson, 2006).

Proper factorisation of matrix is a time consuming process because it needs hundreds of iterations. Numbers of computations in single iteration depend upon the number of input samples. By ignoring the data which consists of negligible signal energy, however, the number of samples for operation can be reduced. Although an audio signal has frequencies in the range of 20–20 000 Hz, most of the information of speech signal is contained in the lower frequencies. In practice, higher frequencies are mostly affected by real time random noise (generated by recording camera and environment), therefore, by processing only lower frequency samples, faster operation may be achieved without any noticeable degradation in the quality of separation.

Initially, sparse NMF was used to separate mixed images. Hoyer (2004) reported successful implementation of NMF with sparse constraints for image separation. The use of NMF for speech application and its advantages over Independent Component Analysis (ICA) are reported in (Cho *et al.,* 2003). Benetos *et al.* (2006) used sparse NMF (SNMF) application for classification of individual musical instrument from a mixed sound. Demir *et al.* (2013) have shown the demixing of music material made by jingle catalog and speech using NMF.

Single channel speech separation using sparse NMF was proposed by Schmidt and Olsson (2006). Recently, Wang *et al.* (2014) have shown that the perfor-

mance may be improved with suitable choice of basis vectors. In the present work a SNMF is used along with rejection of data which contains negligible energy to separate two individual signals from a single channel mixed signal with low processing time.

Correlation between mixed speech signal and original signal, SDR, signal to artifacts ratio (SAR), and SIR between separated signals are found in order to compare the results of existing and proposed algorithm. In present work it is found that the correlation between separated and original signals and SDR and SIR of separated signals is increased as desired. Here, it is reported that the proposed model is performing better to separate mixed signal than an existing algorithm based on these parameters. The proposed model is also showing faster operation, which makes it applicable for real time application.

The paper is organised as follows. In Sec. 2, a brief review of NMF has been given. In Sec. 3, sparse NMF with the criterion of choosing of sparse parameter is elaborated. Designing of model with the help of NMF and frequency selection using wavelet transform is discussed in Sec. 4 followed by results and conclusion.

2. Non-negative matrix factorisation

Consider two individual speech source generating signals $s_1(t)$ and $s_2(t)$. A microphone is capturing signal which is a mixture of individual signals as:

$$s(t) = p * s_1(t) + q * s_2(t), \tag{1}$$

where p and q are the scaling factor by which individual speech signals are affected, which depends upon the distance of speakers from the microphone. An NMF factorises the spectrogram magnitude ($\mathbf{X}$) of the mixed speech signal and has N column vectors of length M. M depends on the frequency resolution taken at the time of Short Time Fourier Transform (STFT), whereas N depends upon the length of speech signal. All elements of the matrix $\mathbf{X}$ are non-negative. The NMF factorises $\mathbf{X}$ as follows:

$$\mathbf{X} = [W]\,[H], \tag{2}$$

where the basis matrix $\mathbf{W}$ is of $M \times K$ order and weight matrix $\mathbf{H}$ is of $K \times N$. K is the number of basis vectors for $\mathbf{X}$. The size of K should be less or equal to min (M, N).

To find out closest factorisation, a cost function between $\mathbf{U}$ and $\mathbf{V}$ ($\mathbf{U}$ is original spectrogram magnitude matrix and $\mathbf{V}$ is reconstructed spectrogram magnitude matrix) is defined in terms of Euclidean distance, i.e., $\sum_{i,j}(U_{i,j} - V_{i,j})^2$, however, for speech/audio applications Itakura-Satio or Kullback-Leibler (K-L) divergence is often found to be more suitable (Nasersharif, Abdali, 2015; Févotte *et al.*, 2009; Lee, Seung, 2000). K-L divergence is defined as:

$$D(\mathbf{U}\|\mathbf{V}) = \sum_{ij}\left(U_{ij}\log\frac{U_{ij}}{V_{ij}} - U_{ij} + V_{ij}\right). \tag{3}$$

As $\mathbf{U}$ and $\mathbf{V}$ are not symmetric, so $D(\mathbf{U}\|\mathbf{V})$ is not termed as Euclidean distance but it is also lower bound by zero at $\mathbf{U} = \mathbf{V}$. For better approximation of factorisation of signal $\mathbf{X}$ (Eq. (2)), $D(\mathbf{X}\|\mathbf{WH})$ should be minimised (Zhu *et al.*, 2013). Initially, random values of $M \times K$ dimensions are assigned to $\mathbf{W}$, and $\mathbf{H}$ is calculated accordingly. This random initialisation affects the quality of factorisation. Better initialisation of basis vector may lead to better approximation. Divergence may be minimised by updating either basis vectors $\mathbf{W}$ or their weight $\mathbf{H}$, or both $\mathbf{W}$ and $\mathbf{H}$. For finding basis vectors of individual training signals, update in both $\mathbf{W}$ and $\mathbf{H}$ is required. This will provide most appropriate basis vectors. Consider spectrogram magnitude of signals $s_1(t)$ and $s_2(t)$ as $\mathbf{X}_1$ and $\mathbf{X}_2$. Then $\mathbf{X}_1$ and $\mathbf{X}_2$ may be given by:

$$\mathbf{X}_1 = [W_1]\,[H_1]\,, \qquad \mathbf{X}_2 = [W_2][H_2]. \tag{4}$$

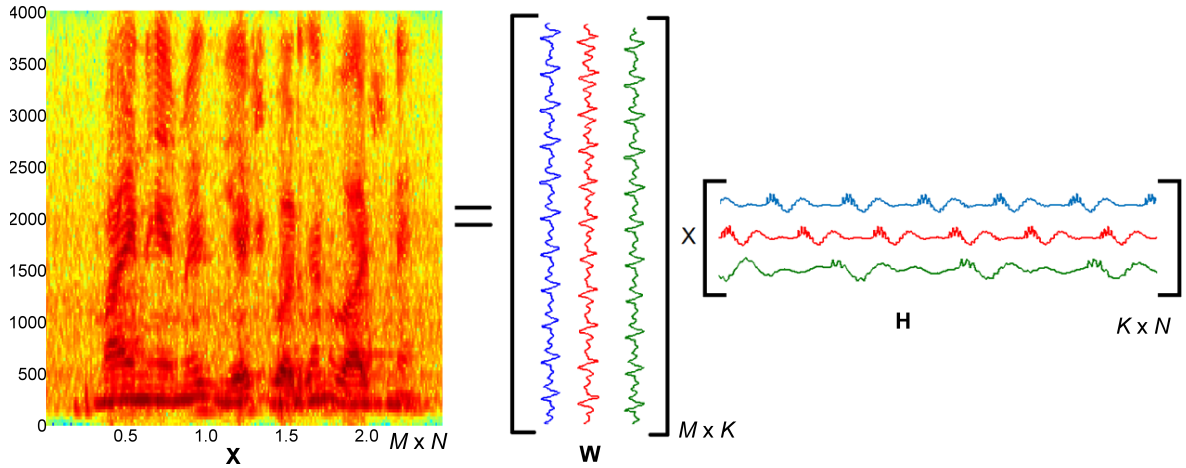

Fig. 1. Non-negative matrix factorisation of a speech signal represented by three basis vectors.

By concatenating basis vectors of both the signals, a weight matrix for the mixed signal is found as:

$$[W] = [W_1\, W_2]. \tag{5}$$

According to the generated basis vector of the mixed signal, the weight matrix will be calculated using SNMF. The source separation from extracted **W** and **H** matrix is done by separating signal portions as shown here:

$$\mathbf{X} = [W_1\, W_2] \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} = \mathbf{X}_1 + \mathbf{X}_2. \tag{6}$$

Number of basis vectors K also affects the performance of NMF. A small value of K will result in greater error because limited number of basis vector may not be able to represent the original signal. For large value of K, the signal extraction does not improve much, while the computational time increases drastically. Wang, Sha (2014) reported the effect of number of basis vectors used for defining any signal.

3. Sparse NMF

One major drawback of conventional NMF is its inability to use the sparseness between different speech signals. Actually conventional NMF is not bothered about the sparseness of individual signals, which reduces the quality of separation. As speech signals are having sparse characteristics, so it will have sparse representation of data. By adding sparseness constraint in to NMF, controllability over sparse representation of output can be extended (Wang, Sha, 2014; Kim, Park, 2008). Sparsity can be imposed on the weight matrix with the help of sparse parameter. Here sparse parameter is represented by γ.

The divergence formulation stated in Eq. (3) is modified into (7) as:

$$\min_{\mathbf{W},\mathbf{H}} D\left(\mathbf{X}\|\mathbf{WH}\right) = \min_{\mathbf{W},\mathbf{H}} \left[\|\mathbf{X}-\mathbf{WH}\|_F^2 + \gamma \sum_{i,j} H_{i,j} \right], \tag{7}$$

$$\mathbf{W},\mathbf{H} \geq 0,$$

where $\|.\|_F$ is denoting the Frobenius norm. The sparse parameter affects the weight matrix **H** and basis vector **W** as described in Eqs. (8) and (9) respectively:

$$H_{i,j} \leftarrow H_{i,j} * \frac{X_i^T \overline{W}_j}{[WH]_i^T \overline{D}_j + \gamma}, \tag{8}$$

$$W_j \leftarrow W_j * \frac{\sum_i H_{i,j} \left[X_i + ([WH]_i^T \overline{W}_j) \overline{W}_j \right]}{\sum_i H_{i,j} \left[[WH]_i + (X_i^T \overline{W}_j) \overline{W}_j \right]}. \tag{9}$$

The sparse parameter γ is chosen as larger when stronger sparsity exists but it leads to relatively poor approximation, while smaller values of γ can be used for better accuracy of approximation but the number of iterations to reach the minima of cost function is increased. Time taken for processing the signal can also be managed by choosing a proper sparse parameter (Hoyer, 2004; Kim, Park, 2008).

4. Frequency selection based speech separation

After analysing approximately 240 speech signals of 2 to 3 seconds from 30 different speakers (both males and females) with the sampling frequency of 16 kHz and 48 kHz (wideband microphone speech signal and standard audio signal from digital video equipment), it has been found that more than 95% of signal energy is contained in the lower 50% frequency band of signal. For reference a speech signal of English digit 'one' 'two' 'three' 'four' sampled at 16 kHz and 48 kHz and their wavelet decompositions are shown in Fig. 2.

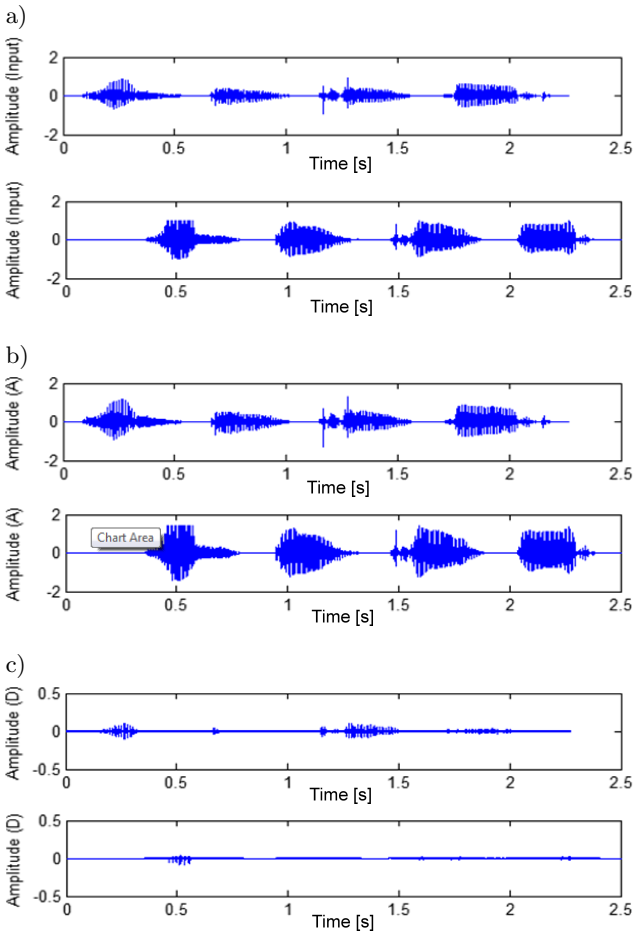

Fig. 2. Speech signals plot for digits 'one' 'two' 'three' 'four' sampled at 16 and 48 kHz, respectively: a) input signals, b) signals containing low frequencies from 0–4 kHz and 0–12 kHz, c) signals containing high frequencies from 4–8 kHz and 12–24 kHz, respectively.

Figure 2a shows the original signals sampled at 16 kHz and 48 kHz in 1st and 2nd row. Signal of lower

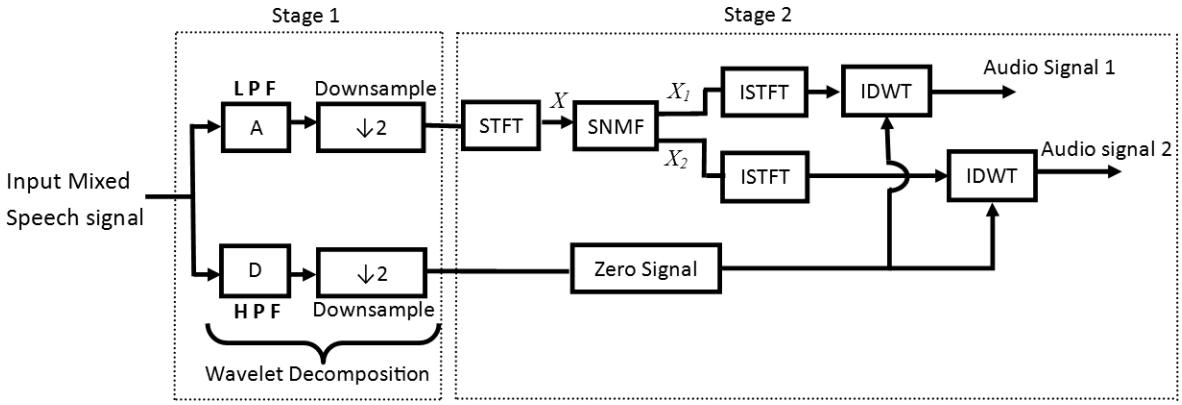

Fig. 3. Proposed model for speech separation.

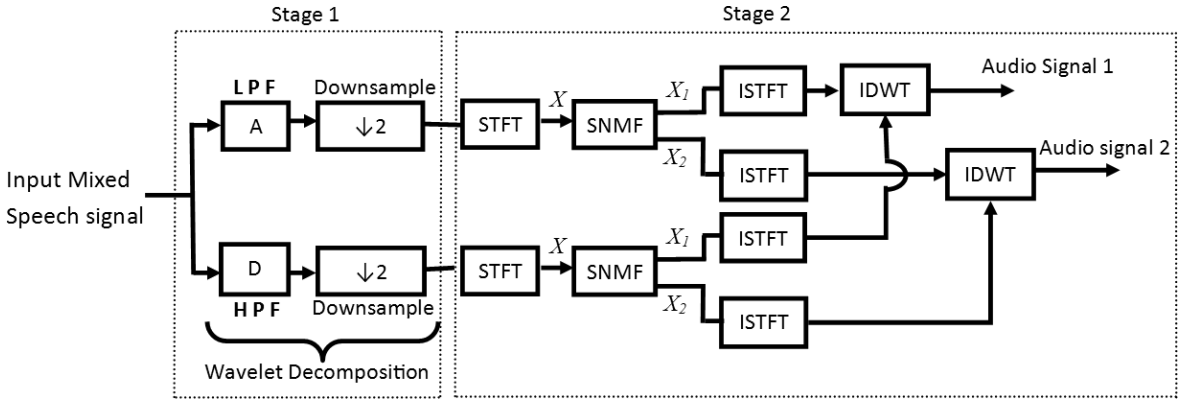

Fig. 4. High frequency signal separation model.

half frequencies denoted by 'A' and signal of higher half frequencies 'D' is shown in Fig. 2b and Fig. 2c, respectively.

The proposed speech separation model is a cascaded structure. The first stage of the model is a system based on wavelet decomposition. Second stage separates the speech signal using SNMF. The frequency components containing higher amount of signal energy are sent for further processing, whereas the remaining signal components are kept for the reconstruction of signal at output as shown in Fig. 3.

To reduce the size of signal for separation, signals are filtered by wavelet decomposition consisting of a low pass and a high pass filters followed by down sampler. A signal coming from lower 50% of frequency band is sent for factorisation. A high frequency signal is considered as noise and replaced by zero of the same length. Spectrograms of low frequency signals are obtained by short time Fourier transforms (STFT). Sparse NMFs (SNMFs) factorize the spectrogram magnitude matrix $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$. After factorisation, speech signals are obtained from $\mathbf{X}_1$ and $\mathbf{X}_2$ by inverse STFT (ISTFT). Separated speech signals s_1 and s_2 are reconstructed by inverse wavelets.

In this paper, high frequency signal separation is also tried in place of zero replacement for high frequency speech signals as shown in Fig. 4. The separated high frequency signals are recombined with low frequency separated signals using IDWT.

5. Evaluation parameters

5.1. Correlation value

The correlation value of mixed and separated signals with individual signals (1st and 2nd signals) has been calculated by using the following expression (Walpole *et al.*, 2011):

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}, \quad (10)$$

where r is the sample correlation coefficient, n is the sample size, x is the value of the 1st variable, y is the value of the 2nd variable.

5.2. Global performance measures of source separation

The common distortion measures SDR, SIR, SAR are described in Eqs. (11)–(13), see (Vincent *et al.*,

2006) by using (FÉVOTTE *et al.*, 2005). The parameters are defined as:

$$\text{SDR} \triangleq 10\log_{10}\frac{\|s_t\|^2}{\|e_{\text{interf}}+e_{\text{artif}}\|^2}, \tag{11}$$

$$\text{SIR} \triangleq 10\log_{10}\frac{\|s_t\|^2}{\|e_{\text{interf}}\|^2}, \tag{12}$$

$$\text{SAR} \triangleq 10\log_{10}\frac{\|s_t+e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2}, \tag{13}$$

where $\widehat{s}_i = s_t + e_{\text{interf}} + e_{\text{artif}}$ is the estimated/reconstructed signal, $s_t \triangleq \langle \widehat{s}_i, s_i\rangle s_i$ is the targeted source, $e_{\text{interf}} \triangleq \langle \widehat{s}_i, s_{i'}\rangle s_{i'}$ is the interference error, $e_{\text{artif}} \triangleq \widehat{s}_i - (s_t + \langle \widehat{s}_i, s_{i'}\rangle s_{i'})$ is the artifacts error, $s_{i'}$ are the input signals other than s_i, $\langle a, b\rangle := \sum_{t=0}^{T-1} a(t)\overline{b}(t)$ is the inner product between two signals a and b of the same length and $\overline{b}$ is a complex conjugate of b.

5.3. Computational time

The implementation of the proposed method was carried out on MATLAB 2011a working on 64 bit Windows 7 operating system running on a 1.9 GHz Intel i7 processor with 4 GB RAM. The computation for separation time of mixed speech data is done. Time consumed to find the basis vectors from training signal is not considered because it is assumed that these vectors will be recorded already and then any source separation technique will be applied.

6. Simulations and results

To evaluate the performance of the proposed approach, simulations have been performed on speech signals from TIMIT database which contains wideband microphone speech signals sampled at 16 kHz. Further the speech signals of digital video equipment sampled at 48 kHz are taken for simulation from Aligarh Muslim University Audio Data Library (AMUADLib) (UPADHYAYA *et al.*, 2013). AMUADLib contains 2 common and 8 different sentences by 100 speakers of both genders. 400 and 2368 combinations of male-female, 180 and 1856 combinations of female-female and 180 and 4228 combinations of male-male speakers of 16 kHz and 48 kHz sampled sentences are taken for analysis of models, respectively. The average length of individual speech signals taken for the experiment is 2.1 seconds from TIMIT data and 2.55 seconds from AMUADLib. As each sentence has a different length so zero padding is applied for addition of both signals.

From AMUADLib, Hindi and English digits data ('Ek' 'Do' 'Teen' 'Char' 'Paanch' 'Chai' 'Saat' 'Aath' 'Now' 'Dus' and 'one' 'two' 'three' 'four' 'five' 'six' 'seven' 'eight' 'nine' 'ten') have been taken for training purpose from each speaker. From TIMIT database eight sentences have been taken for training purpose from each speaker. The average length of a training signal is about 8.8 seconds. 512 point Fast Fourier Transform with 50% overlap was used to find STFT with window size of 10 milliseconds. Based on these training signals, basis vectors are found using SNMF as described in Sec. 3. Here the number of basis vector taken is $K = 100$. The sparseness parameter is fixed for our simulation purpose, i.e., 0.5, which is enough to separate sparse signals with a moderate computational time.

A speech signal mixture of male-male (MM), male-female (MF) and female-female (FF) has been separated with and without using the frequency based selection (Wavelet Decomposition). Table 1 can be used for comparing the overall performance of the existing model with the proposed model for 16 kHz. Table 1 also contains the results for the model in which the high frequency signals are separated and the individual signals reconstructed using IDWT with separated low frequency signals.

Table 1. Performance of the existing model and proposed model for separation of mixed signal for 16 kHz sampled signal.

Gender	Separation time [s]	Corr_1st_mix	Corr_2nd_mix	Corr_1st_x1	Corr_2nd_x2	SDR [dB]	SIR [dB]	SAR [dB]
Existing Model (without using Wavelet Decomposition)								
MM	0.89416	0.65814	0.72128	0.71074	0.766438	1.912557	3.654271	9.003129
FF	0.95930	0.69155	0.70432	0.76841	0.783969	2.867325	5.249127	8.435939
MF	0.92917	0.62508	0.75723	0.82236	0.870187	5.266586	8.906659	8.672519
Proposed Model (using Wavelet Decomposition)								
MM	0.47996	0.65814	0.72128	0.72100	0.784089	2.464281	5.316304	7.535073
FF	0.49931	0.69155	0.70432	0.76401	0.768281	2.9944	6.211138	7.537197
MF	0.48005	0.62508	0.75723	0.83347	0.866885	5.792528	10.80475	8.193429
Signal Separation of Both Lower and Higher Frequency Components								
MM	1.28554	0.65814	0.72128	0.70301	0.75841	1.890813	5.233115	7.620994
FF	0.90763	0.69156	0.70433	0.76826	0.78443	3.058805	5.978104	7.848247
MF	0.86949	0.62509	0.75723	0.83382	0.87869	5.79779	10.46752	8.383128

Separation time is the time required to separate mixed signals using the given algorithm in seconds. Corr_1st_mix and Corr_2nd_mix are the correlation coefficients of the first speaker's signal with mixed speech signal and second speaker's signal with mixed speech signal, respectively. After the separation of signals at the output end, Corr_1st_x1 and Corr_2nd_x2 are calculated to show the correlation coefficients of the first speaker's input signal with the first separated signal and the second speaker's input signal with the second separated signal, respectively. All results are the average of 5 random initialisations.

The following histogram shows the average correlation improvement in percentage using the existing and proposed algorithms (Fig. 5).

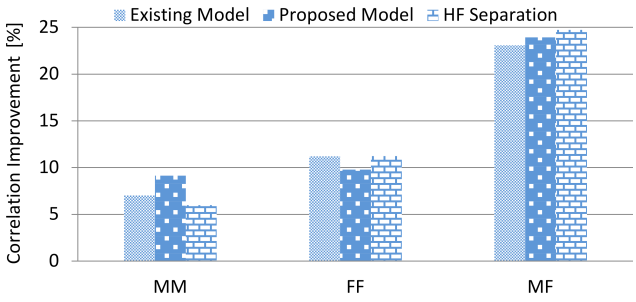

Fig. 5. Correlation improvement in percentage of individual speech signals with mixed signal and separated signals using the existing and proposed models.

For all cases, 13.76% and 14.27% average improvement is found in correlation between the original and separated signals using the existing and proposed models, respectively. Correlation improvement by high frequency signal separation model is 13.97%, which is better than the existing algorithm but not better than the proposed model.

Figure 6 shows the improvement in SDR and SIR using the proposed model in comparison with the existing algorithm. The results matched to our expectations in terms of SDR and SIR. However, the proposed algorithm leads to some artifacts result in lower SAR. The higher SIR and lower SAR are due to the lower projection of the reconstructed signal to the other signals than the original individual required signal $\langle \hat{s}_i,\ s_{i'} \rangle$. A lower projection leads to a low interference but higher artifacts error.

SDR of the same gender after separation is 2.38 dB and 5.27 dB for the opposite gender using the existing model, where the proposed model shows 2.72 dB SDR for the same gender and 5.79 dB for the opposite gender. SIR is also improved from 4.43 dB to 5.76 dB for the same gender and 8.90 dB to 10.80 dB for the opposite gender. But SAR is reduced from 8.72 dB to 7.53 dB for the same gender and 8.67 dB to 8.19 dB for the opposite gender. For high frequency separation models, SDR for the same and opposite genders

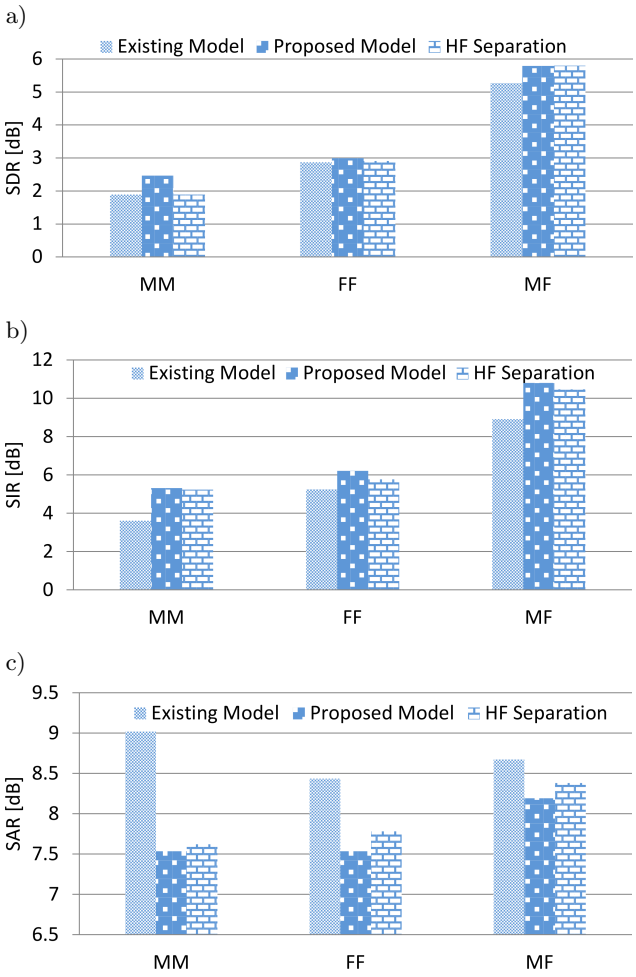

Fig. 6. Comparative performance evaluation of the existing and proposed models using: a) signal to distortion ratio, b) signal to interference ratio, c) signal to artifacts ratio.

is 2.40 dB and 5.80 dB, respectively. SIR for the same and opposite genders is 5.50 dB and 10.46 dB, respectively. And SAR for the same and opposite genders is 7.70 dB and 8.38 dB, respectively.

As the basis vectors of training signals considered in the data set, the time taken for extraction of the basis vector is not considered in the calculation. However, all the above results may change slightly in every experiment, as these are highly dependent upon the initialisation of the basis vector taken by the system.

As mentioned earlier, the test signals have the average time duration of 2.55 seconds. It takes for the existing model about 0.92 seconds to separate two signals using the given system configuration, whereas it takes around 0.49 seconds for the proposed model and around 0.91 seconds for the high frequency signal separation model, as shown in Fig 7. The proposed model reduces the time requirement to separate signals by approximately 46.73%.

All results show that the high frequency signal separation model is better than the existing model but the

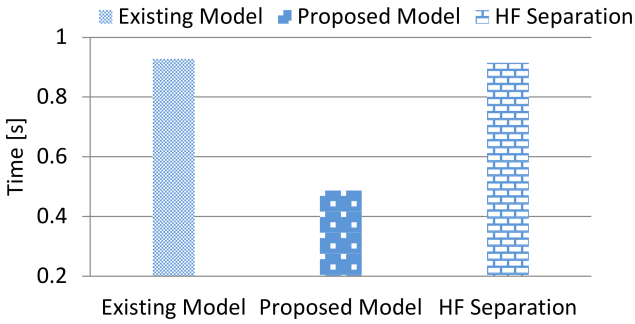

Fig. 7. Computational time to separate signals from the mixed signal using the existing and proposed models.

results are not so remarkable with negligible improvement in the computational burden. So, further simulations for the mixed speech signal sampled at 48 kHz are performed on the existing and proposed models only. Table 2 can be used for comparing the overall performance of the existing model with the proposed one for 48 kHz.

Figure 8 shows the average correlation improvement in percentage using the existing and proposed algorithms for speech signals sampled at 48 kHz.

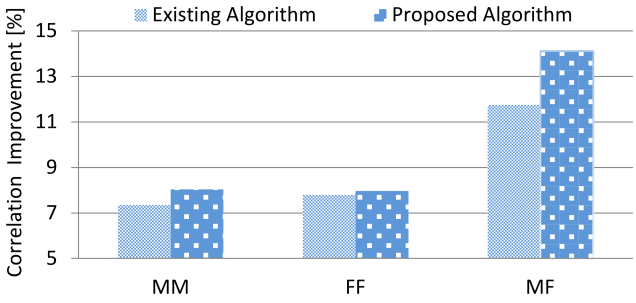

Fig. 8. Correlation improvement in percentage of individual speech signals with mix signal and separated signals using the existing and proposed models.

For all cases, 8.96% and 10.23% average improvement is found in correlation between the original and separated signals using the existing and proposed models, respectively. It can be said that the proposed algorithm performs 1.23% better than the existing one.

Figure 9 shows the improvement in SDR and SIR using the proposed model in comparison with the existing algorithm.

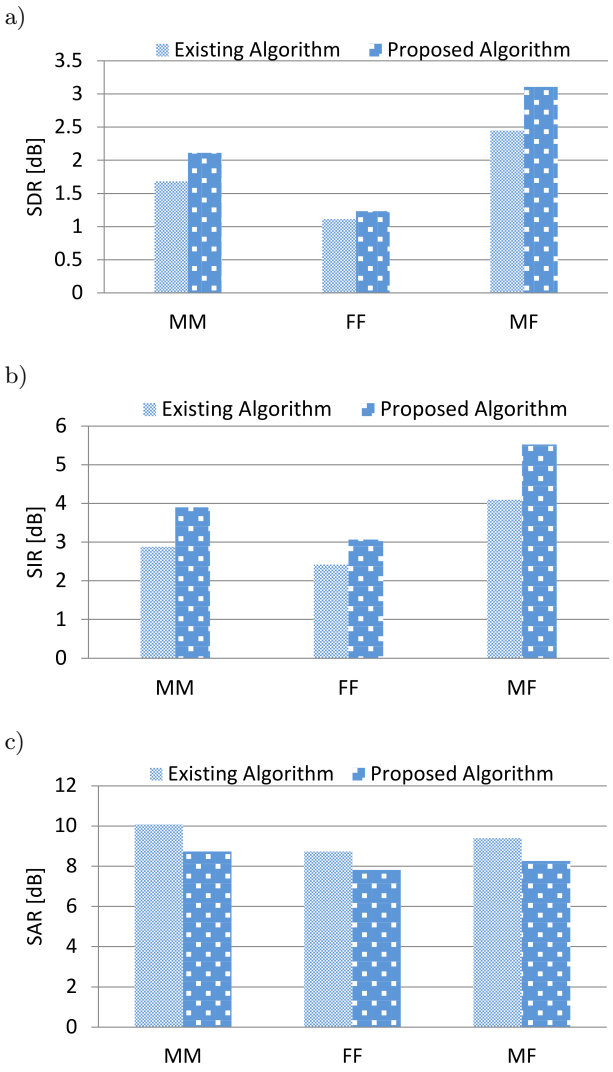

Fig. 9. Comparative performance evaluation of the existing and proposed models using: a) signal to distortion ratio, b) signal to interference ratio, c) signal to artifacts ratio.

Table 2. Performance of existing model and proposed model for separation of mixed signal for 48 kHz sampled signal.

Gender	Separation time [s]	Corr_1st_mix	Corr_2nd_mix	Corr_1st_x1	Corr_2nd_x2	SDR [dB]	SIR [dB]	SAR [dB]
Existing Model (without using Wavelet Decomposition)								
MM	3.24725	0.70144	0.68516	0.75158	0.73708	1.68318	2.88217	10.0829
FF	3.26976	0.70009	0.70061	0.75468	0.75515	1.11444	2.41860	8.73423
MF	3.47564	0.70729	0.68841	0.79254	0.76728	2.44976	4.09292	9.39623
Proposed Model (using Wavelet Decomposition)								
MM	1.61918	0.70144	0.68516	0.75821	0.73992	2.11094	3.90159	8.73423
FF	1.63695	0.70009	0.70061	0.75364	0.75872	1.23419	3.06655	7.80943
MF	1.74793	0.70729	0.68841	0.81369	0.77920	3.10664	5.52209	8.25576

SDR of the same gender after separation is 1.4 dB and for the opposite gender it is 2.44 dB using the existing model, where the proposed model shows 1.67 dB SDR for the same gender and 3.10 dB for the opposite gender. SIR is also improved from 2.65 dB to 3.48 dB for the same gender and 4.09 dB to 5.52 dB for the opposite gender. But SAR is reduced from 9.4 dB to 8.26 dB.

As mentioned earlier, the test signals have the average time duration of 2.55 seconds. It takes about 3.33 seconds for the existing model to separate two signals using the given system configuration, whereas it takes around 1.67 seconds for the proposed model, as shown in Fig. 10. The proposed model reduces the time requirement to separate signals by approximately 49.89%. This result may lead to real time speech separation using small packets of mixed speech signal.

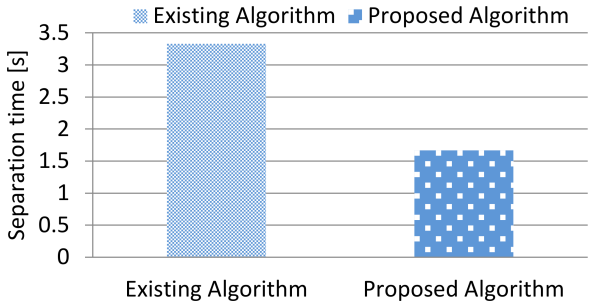

Fig. 10. Computational time to separate signals from the mixed signal using the existing and proposed models.

Performance of the algorithm also depends upon the intensity (loudness) of the speech signal and difference in formant frequencies of speakers. If the intensity of speech signals by two speakers is very different then there may be a problem of masking, which results in poor separation of speech signals. Moreover, if formant frequencies of the speech signals are different then the quality of speech signals separation will be better.

As the energy content of speech signal is more in 1st and 2nd formant frequencies (Reetz, Jongman, 2011), the relationship between the difference of 1st and 2nd formant frequency (f_{d_1} and f_{d_2}) of two individual speech signals and the average improvement in terms of correlation of individual signals from the experiment is reported in Fig. 11, where f_{d_1} and f_{d_2} are calculated as:

f_{d_1} = 1st formant frequency of one speech signal – 1st formant frequency of other speech signal,

f_{d_2} = 2nd formant frequency of one speech signal – 2nd formant frequency of other speech signal.

This shows that the average improvement in correlation between separated and original signals increases

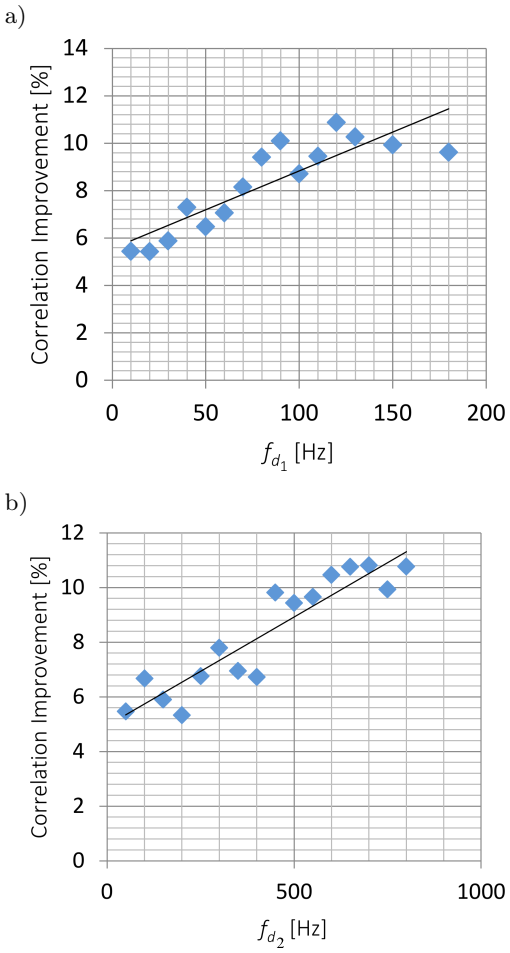

Fig. 11. Difference of (a) 1st and (b) 2nd formant frequencies of individual signals v/s correlation improvement after separation.

with the increment in difference of 1st and 2nd formant frequencies of both individual original signals, respectively.

7. Conclusion

Noise generated by microphone at the time of recording contains high frequencies components. By removing them using the wavelet decomposition, separation of mixed speech signal is done with improvement in performance over the existing algorithm in terms of correlation, SDR, and SIR. This also results in a lower execution time of the algorithm. The effect of formant frequencies over separation capability of the proposed model is also shown. It is also reported that for better separation of the mixed signal, the intensity of speakers should be nearby equal and the formant frequencies should have enough difference.

However, a lot of improvisation can be carried out to improve the performance and speed of separation. Better initialisation leads to earlier optimization, which reduces the number of iterations to optimise the

cost function and better separation. The selection of the proper sparse parameter according to speech signals which are mixed together is the issue to solve. It may also play an important role in proper and fast finding of the weight matrix.

References

1. Benetos E., Kotti M., Kotropoulos C. (2006), *Musical instrument classification using non-negative Matrix factorization algorithms and subset feature selection*, IEEE International Conference on Acoustics, Speech and Signal Processing, **5**, 221–224.
2. Cho Y-C., Choi S., Bang S-Y. (2003), *Non-negative component parts of sound for classification*, 3rd IEEE International Symposium on Signal Processing and Information Technology, 633–636.
3. Demir C., Saraclar M., Cemgil A.T. (2013), *Single-channel speech-music separation for robust ASR with mixture models*, IEEE Transactions on Audio, Speech, and Language Processing, **21**, 4, 725–736.
4. Févotte C., Bertin N., Durrieu J. (2009), *Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis*, Neural Computation, **21**, 793–830.
5. Févotte C., Gribonval R., Vincent E. (2005), *BSS_EVAL toolbox user guide revision 2.0,* Tech. Rep. 1706, IRISA, Rennes, France.
6. Hoyer P.O. (2004), *Non-negative matrix factorization with sparseness constraint*, Journal of Machine Learning Research, 1457–1469.
7. Kim J., Park H. (2008), *Sparse nonnegative matrix factorization for clustering*, Georgia Institute of Technology, GT-CSE-08-01.
8. Lee D.D., Seung H.S. (1999), *Learning the pans of objects with nonnegative matrix factorization*, Nature **401**, 788–791.
9. Lee D.D., Seung H.S. (2000), *Algorithms for nonnegative matrix factorization*, Advances in Neural Information Processing Systems, **13**, 556–562.
10. Nasersharif B., Abdali S. (2015), *Speech/music separation using non-negative matrix factorization with combination of cost functions*, International Symposium on Artificial Intelligence and Signal Processing (AISP), 107–111.
11. Paatero P., Tapper U. (1994), *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics, **5**, 111–126.
12. Reetz H., Jongman A. (2011), *Phonetics: transcription, production, acoustics, and perception*, Wiley-Blackwell, ISBN: 978-1-4443-5854-4, pp. 182–200.
13. Schmidt M.N., Olsson R.K. (2006), *Single-channel speech separation using sparse non-negative matrix factorization*, 9th International Conference on Spoken Language Processing, Pittsburgh, PA, USA.
14. Upadhyaya P., Farooq O., Varshney P., Upadhyaya A. (2013), *Enhancement of VSR using low dimension visual feature*, International Conference of Multimedia, Signal Processing and Communication Technologies (IMPACT), Aligarh, India, pp. 71–74.
15. Vincent E., Gribonval R., Fevotte C. (2006), *Performance measurement in blind audio source separation*, IEEE Transactions on Audio, Speech, and Language Processing, **14**, 1462–1469.
16. Walpole R.E., Myers R.H., Myers S.L., Ye K.E. (2011), *Probability and Statistics for Engineers and Scientists*, 9th ed., Pearson, ISBN: 978-0-3216-2911-1, p. 433.
17. Wang Y., Li Y., Ho K.C., Zare A., Skubic M. (2014), *Sparsity promoted non-negative matrix factorization for source separation and detection*, 19th International Conference on Digital Signal Processing (DSP), 640–645.
18. Wang Z., Sha F. (2014), *Discriminative non-negative matrix factorization for single-channel speech separation*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3749–3753, Florence, Italy, 4–9 May.
19. Zhu B., Li W., Li R., Xue X. (2013), *Multi-stage non-negative matrix factorization for monaural singing voice separation*, IEEE Transactions on Audio, Speech, and Language Processing, **21**, 10, 2096–2107.