Paper

Model-Based Method for Acoustic Echo Cancelation and Near-End Speaker Extraction: Non-negative Matrix Factorization

Pallavi Agrawal and Madhu Shandilya

Department of Electronics and Communication Engineering,
Maulana Azad National Institute of Technology, Bhopal, India

https://doi.org/10.26636/jtit.2018.122617

Abstract—Rapid escalation of wireless communication and hands-free telephony creates a problem with acoustic echo in full-duplex communication applications. In this paper a simulation of model-based acoustic echo cancelation and near-end speaker extraction using statistical methods relying on non-negative matrix factorization (NMF) is proposed. Acoustic echo cancelation using the NMF algorithm is developed and its implementation is presented, along with all positive, real time elements and factorization techniques. Experimental results are compared against the widely used existing adaptive algorithms which have a disadvantage in terms of long impulse response, increased computational load and wrong convergence due to change in near-end enclosure. All these shortcomings have been eliminated in the statistical method of NMF that reduces echo and enhances audio signal processing.

Keywords—adaptive algorithms, convergence, echo cancelation, non-negative matrix factorization (NMF).

1. Introduction

In the era of wireless communication, widespread use of hands-free telephony has been observed. This results in disturbances from acoustic echo and associated noise, which decreases the quality of speech [1]. Echo is a phenomenon in which a delayed and distorted original signal is reflected to its source. Echo in audio speech occurs when the sound is reflected from nearby objects, walls or from the floor. If these reflections are of a short duration and arrive in a very short time, they are referred to as [2] reverberations or the spectral distortion. Whereas if the same sound arrives back within a few tens of milliseconds, it is heard as a distinct reflected sound known as acoustic echo. In a telephony system, two types of echo may be distinguished: network echo [3], [4] and acoustic echo [5]–[7]. Network echo is mostly created along telephone lines due to an impedance mismatch between public switched telephone networks. The phenomenon of acoustic echo occurs mostly in hands free communications. Earlier work in the area of echo cancelation (EC) focused primarily on network echo cancelation. With advances in wireless communication technologies, cancelation of acoustic echo has captured attention of users. Figure 1 shows a scenario of acoustic echo affecting a teleconferencing system.

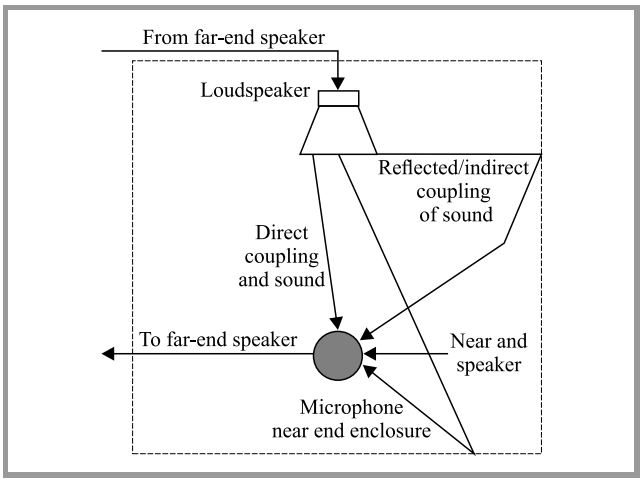

Fig. 1. Generation of acoustic echo.

Acoustic echo [8] occurs when audio from the far-end speaker comes arrives at the near-end enclosure via a loudspeaker and is picked up by the near-end microphone via both direct and indirect paths. To remove the echo, echo cancelers are developed which detect and remove the echo generated. In order to calculate the adaptive filter tap, various algorithms, such as least mean squares (LMS), normalized LMS and recursive LMS are employed. The main features of these algorithms are that they offer fast convergence, but at the cost of computational complexity observed with an increase of the number of filter taps. Also, any change in near-end enclosure may lead to wrong convergence, and additional filters may then be required for its removal [9]–[11].

Acoustic echo cancelation (AEC) and extraction of near-end signal is a challenging process, as the proposed method should emulate the inherent ability of the human auditory system, known as computational auditory scene analysis (CASA) [12]–[14].

In this paper, we extend the concept from the conventional method, to model-based statistical pattern recognition. This is motivated by the fact that in the spectral domain, distinct speakers have distinct patterns. To separate these features, machine learning and matrix factorization methods are applied. The main theme is to use, in advance, prior information from the sources, and to first train the model for each source, and then separate signals from a given mixture. Hence, such an approach is known as model-based near-end speaker extraction and echo cancelation. Model-based source separation can be a probabilistic model, or a matrix factorization-based method. The former method uses the Bayesian approach, while the latter uses such techniques as non-negative matrix factorization (NMF). In this paper, model-based methods of AEC, relying on matrix factorization-based methods, are proposed. It is shown that the proposed method outperforms the classic echo cancelation method.

The paper is organized as follows: Section 2 is devoted to problem formulation and contains a brief description of LMS, NLMS, RLS and NMF methods. The model-based AEC and near-end speaker extraction method is presented in Section 3. Algorithmic steps and their implementation are shown in Section 4. Performance analysis and experimental results are shown in Sections 5–6. Simulation results are discussed in Section 7. Section 8 summarizes the paper.

2. AEC and Near-End Speaker Extraction

The process of generating acoustic echo can be described with the use of a linear framework. Let us denote the signal coming from the far-end speaker as $x(n)$, where n is a sample index. Let this excite a linear system whose impulse response is given by $h(n)$. $h(n)$ is actually a superimposition of impulses, with each of them corresponding to an echo. In the near-end enclosure, echo is produced due to various propagation paths between the loudspeaker and the microphone. Let it be represented by $d(n)$, a linear convolution of $x(n)$ with $h(n)$, expressed as:

$$d(n) = \sum_{i=0}^{L_h-1} h(i)x(n-i)\,, \tag{1}$$

where i denotes the sample index and L_h denotes the length of the echo path impulse response. In a vector form, it can be expressed as:

$$\begin{aligned} d(n) &= \mathbf{h}^T\mathbf{x}(n),\ \mathbf{x}(n) \\ &= \left[x(n), x(n-1), \ldots, x(n-L_h+1)\right]^T, \end{aligned} \tag{2}$$

where T denotes the matrix's transpose operation symbol. The output of the near-end microphone signal or the far-end user's signal $y(n)$ is:

$$\begin{aligned} y(n) &= d(n)+u(n)+w(n) = && (3)\\ &= \sum_{i=0}^{L_h-1} h(i)x(n-i)+u(n)+w(n) = && (4)\\ &= \mathbf{h}^T\mathbf{x}(n)+u(n)+w(n)\,, && (5) \end{aligned}$$

where $u(n)$ denotes the desired near-end speech signal and $w(n)$ denote the noise sources (Fig. 2).

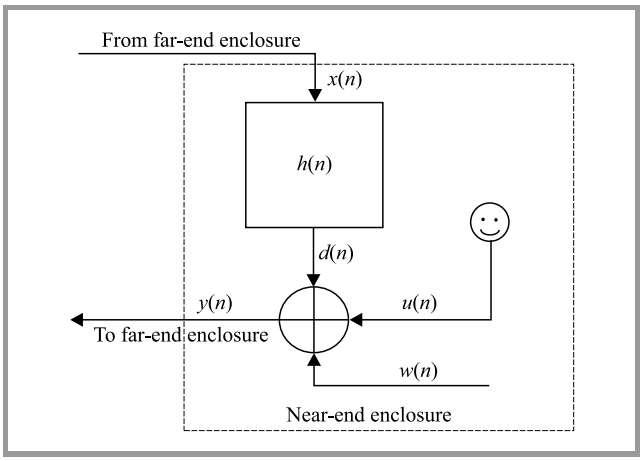

Fig. 2. Linear framework for hands-free communication.

The aim of acoustic echo cancelation is to present a clear near-end speech signal $u(n)$ to the far-end speaker, containing no echo and noise components, i.e. $d(n) = 0$ and $w(n) = 0$.

2.1. LMS Algorithm

The LMS algorithm, derived by Widow and Hoff [15], is one of the most efficient adaptive filtering algorithms. This algorithm has the property of adjusting the coefficients of a filter to reduce MSE between the desired signal and output of the filter. It is used for updating the taps of the adaptive filter during each iteration:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\mathbf{x}^*(n)\,, \tag{6}$$

where $\mathbf{x}(n)$ is the input vector of time-delayed input values, $\mathbf{w}(n)$ is the weight vector at time n, and μ is the step-size parameter that controls the immediate change of the updating factor. Its value has an impact on the performance of the LMS algorithm. When μ is low, it takes a long time for the algorithm to converge and a high value of this factor causes the algorithm to diverge, leading to LMS instability.

2.2. NLMS Algorithm

The step-size parameter used in the LMS algorithm is normalized in the case of normalized least mean squares

(NLMS) [16] algorithm. In NLMS [17], [18], μ for computing the update is given by:

$$\mu(n) = \frac{\beta}{c + \|\mathbf{x}(n)\|^2}\,, \tag{7}$$

where $\mu(n)$ is the step-size parameter at sample n, β is a normalized step-size ($0 < \beta < 2$), and c is the smallest positive constant.

2.3. RLS Algorithm

The Recursive Least Square (RLS) algorithm has a fast convergence rate [30] and is widely used in EC, channel equalization, speech enhancement and radar applications. In this algorithm [25], we consider the following:

- $x(n)$ is the discrete time array $M \times 1$ array input vector,
- $y(n) = w^H x(n)$ is the output signal,
- $d(n)$ is the desired signal,
- w is the $M \times 1$ complex weight matrix.

2.4. NMF Algorithm

Non-Negative Matrix Factorization (NMF) [19] is a linear-based decomposition technique subject to the constraints of non-negativity of the data being decomposed. It actually decomposes the data of a non-negative matrix into two non-negative matrices and a residual matrix which does not necessarily have to be non-negative. A given data matrix $\mathbf{A} \in \mathbb{R}_+^{M\times N}$, it can be decomposed into two non-negative matrices $\mathbf{B} \in \mathbb{R}_+^{M\times K}$ $\mathbf{G} \in \mathbb{R}_+^{K\times N}$, $K < N$ and a residual matrix $E \in \mathbb{R}_+^{M\times N}$ representing the approximation to the error [20], [21].
It is an optimization problem which aims to minimize the cost function $\mathbf{C}(.)$ with respect to $\mathbf{B}$ and $\mathbf{G}$. This cost function measures the divergence between $\mathbf{A}$ and $\mathbf{BG}$.

$$\mathbf{A} = \mathbf{BG} + \mathbf{E},\ \ \mathbf{A} \approx \mathbf{BG}\,. \tag{8}$$

It is an optimization problem which aims at minimizing some cost function $\mathbf{C}(.)$ with respect to $\mathbf{B}$ and $\mathbf{G}$. This cost function measures the divergence between $\mathbf{A}$ and $\mathbf{BG}$.
One of the cost function measures is the Euclidean distance [22]:

$$D_{ED}(\mathbf{A},\mathbf{B},\mathbf{G}) = \frac{1}{2}\left\|\mathbf{A} - \mathbf{BG}\right\|^2. \tag{9}$$

The following multiplicative update rules are followed to balance convergence speed and complexity:

$$b_{ij} \leftarrow b_{ij}\frac{\left[\mathbf{AG}^T\right]_{ij}}{\left[\mathbf{BGG}^T\right]_{ij}},\ \ g_{ij} \leftarrow g_{ij}\frac{\left[\mathbf{B}^T\mathbf{A}\right]_{jk}}{\left[\mathbf{B}^T\mathbf{BG}\right]_{jk}}\,, \tag{10}$$

here $[.]_{ij}$ indicates that the given operations are performed on an element-by-element basis.

The second cost function in use is the generalized version of Kullback-Leibler divergence, also known as I-divergence [23], [24]:

$$D_{KL}(\mathbf{A}\|\mathbf{B},\mathbf{G}) = \sum_{ik}\left(a_{ik}\log\frac{a_{ik}}{[\mathbf{BG}]_{ik}} - v_{ik} + [\mathbf{BG}]_{ik}\right). \tag{11}$$

This cost function is not symmetric in $\mathbf{A}$ and $\mathbf{BG}$. It actually quantifies in bits that how close $\mathbf{A}$ is to $\mathbf{BG}$. Its value equals zero if the distributions match exactly, and infinite if there is no match at all.

3. Model-based AEC and Near-end Speaker Extraction Method

The model-based statistical pattern recognition technique was first proposed in [25]. In the spectral domain there is a distinct pattern of speech signals spoken by different speakers. Due to this regular and distinct pattern of speech signal, matrix factorization methods can be applied to differentiate between these speakers. The NMF [26] approach is formulated in the short-time Fourier transform domain [27], [28]. The near-end microphone signal $y(n)$ of the mixture signal is decomposed into two bases of spectral features. First, training on the magnitude spectra of many speakers is performed offline. This trained data can be called on as and when required. The other bases are created during operation and testing. These bases are continuously updated by the incoming far-end signal $x(n)$ and are actually specific to the far-end signal. Now, NMF is employed that minimizes the cost function. This cost function minimizes divergence of the trained vectors to the test vectors. Once the optimal vector is identified, echo reduction is conducted by performing an inverse transformation for the identified vectors, using the phase information received from the mixture signal.
The model of the acoustic echo which is mostly used in EC and given in the literature [29]–[32] is:

$$|Y(f,k)| = |D(f,k)| + |U(f,k)|\,, \tag{12}$$

where $|Y(f,k)$ is the STFT of $y(n)$, f is the discrete frequency, k is the frame index and $|.|$ is the magnitude of the complex value. For STFT of $y(n)$, hanning window of length N is used that advances in the steps of m. Similarly, $|D(f,k)|$ and $|U(f,k)|$ represent the $d(n)$ and $u(n)$ components of the mixture signal in the STFT domain. Such a model strictly follows non-negativity and linearity of the sample of the speech signals. This allows for easy implementation of NMF for EC.

4. Algorithm Implementation

The pseudocode of the algorithm for NMF-based AEC and near-end speaker extraction approach described above is presented as Algorithm 1. It comprises three stages: training, testing and reconstruction. Training is done for the

Algorithm 1: Algorithm pseudocode for NMF-based AEC and near-end speaker extraction

1. Process the incoming far-end signal $x(n)$
   Find STFT: $X(f,k)$
   $|X(f,k)|$, k is the frame index
   From $X(f,k)$, $d(f,k)$ is created
   Calculate the NMF of $D(f,k)$ and form the basis
   $B_d(k)$ $G_d(k)$
2. Process the speaker independent near-end signal $u(n)$
   Find STFT: $U(f,k)$
   $|U(f,k)|$, k is the frame index
   Calculate the NMF of $U(f,k)$ and form the basis
   $\mathbf{B}_u(k)$ $\mathbf{G}_u(k)$
3. Concatenate the basis of steps 1 and 2 to form composite basis $\mathbf{B}(k) = [\mathbf{B}_u\, \mathbf{B}_d(k)]$
4. Process the mixture signal $y(n)$
   Find STFT: $Y(f,k)$
   $|Y(f,k)|$, k is the frame index
   For $i = 1$ to ϕ (restricted NMF updates) $+\psi$ (unrestricted NMF updates) **do**
   (ϕ and ψ are the number of iterations)
   For ϕ iterations, composite basis $\mathbf{B}(k)$ is fixed and $\mathbf{g}(k)$ is updated as
   $$g(k) \leftarrow g(k)\frac{\mathbf{B}(k)\left[\frac{\mathbf{y}(k)}{\mathbf{B}(k)\mathbf{g}(k)}\right]}{\mathbf{B}(k)^T 1 + \delta},$$ δ is a positive regularization factor
   $y(k)$ can now be expressed as:
   $$\mathbf{y}(k) = \mathbf{B}(k)\mathbf{g}(k) + \mathbf{e}(k) = [\mathbf{B}_u\, \mathbf{B}_d(k)]\begin{bmatrix}\mathbf{g}_u(k)\\ \mathbf{g}_d(k)\end{bmatrix} + \mathbf{e}(k)$$
   For ψ iterations, both composite bases $[\mathbf{B}_u\, \mathbf{B}_d(k)]$ and $\mathbf{g}(k)$ are updated
   $$\mathbf{B}(k) \leftarrow \mathbf{B}(k)\frac{\left[\frac{\mathbf{y}(k)}{\mathbf{B}(k)\mathbf{g}(k)}\right]\mathbf{g}(k)^T}{1\mathbf{g}(k)^T + \delta}$$
   End for
   $\hat{\mathbf{u}}(k) = \mathbf{B}_u\mathbf{g}_u(k)$
5. Near-end speaker extraction
   $\hat{\mathbf{u}}(k) \angle Y(f,k)$

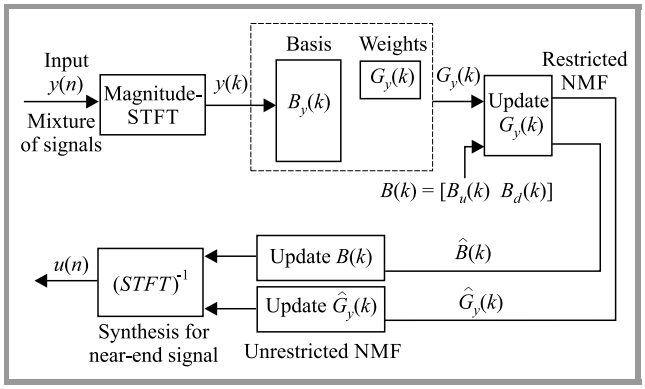

Fig. 3. Block diagram presenting AEC and near-end speaker extraction using NMF.

far-end speaker and for the near-end speakers. From the training phase, the base vectors $\mathbf{B}_d$ and $\mathbf{B}_u$ for far-end and near-end speaker signals are obtained. These bases are concatenated to form composite bases $\mathbf{B} = [\mathbf{B}_u\mathbf{B}_d]$. The composite base is derived during testing. Testing is done online for the incoming mixture signal. A block diagram presenting AEC relying on the NMF algorithm is shown in Fig. 3.

5. Performance Analysis

Spectrograph plot is a three-dimensional (3D) visual analysis of an acoustic signal, in which horizontal axis represents the time domain, the vertical axis represents frequency and

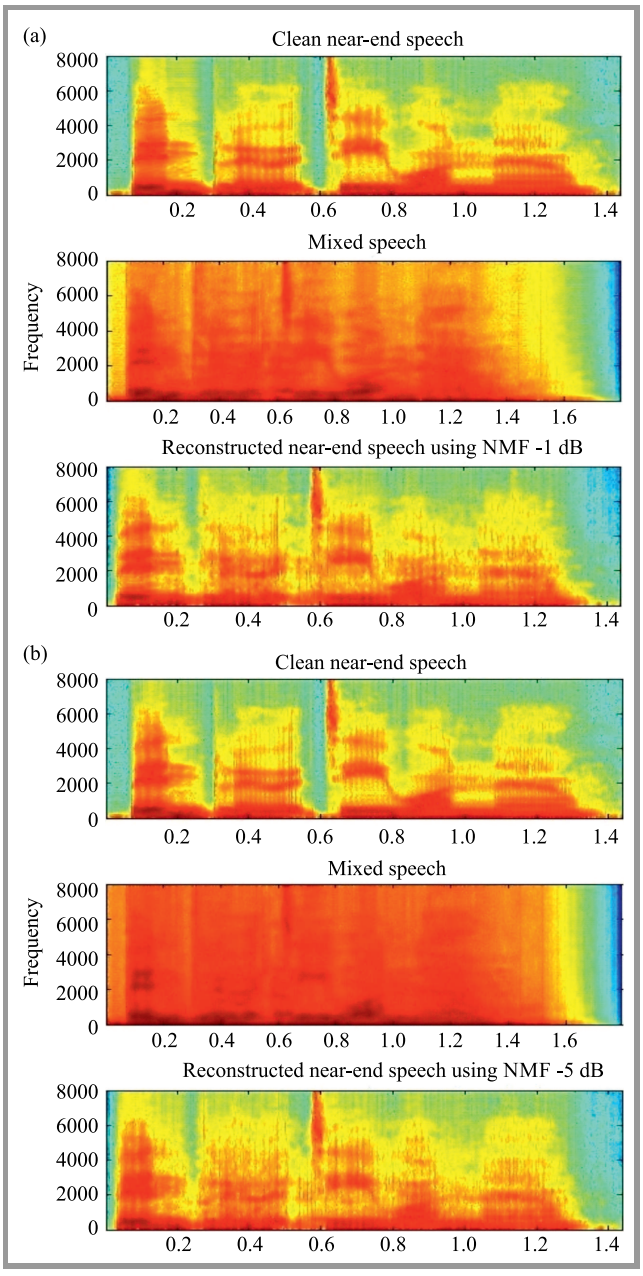

Fig. 4. Spectrogram of extracted near-end signal using NMF at ENR equal to: (a) –1 dB and (b) –5 dB. (For color pictures visit www.nit.eu/publications/journal-jtit)

the third dimension is amplitude at a frequency and time indicated by the color. Figure 4a shows the spectrogram of a near-end signal extracted using the NMF algorithm at ENR –1 dB. Figure 4b shows the spectrogram of a near-end signal extracted using the NMF algorithm at ENR –5 dB. In Figs. 4a-b, the top picture shows the spectrogram of a clean near-end signal, the middle image shows a mixed signal and the bottom picture shows a near-end signal reconstructed using NMF. Comparing the NMF algorithm at ENR –1 dB and ENR –5 dB, one may notice, visually, that a better reconstructed near-end speech signal is seen at ENR –5 dB, compared to ENR of –1 dB. This is due to the reduction in echo from –1 to –5 dB, and, hence, to better near-end speaker extraction.

5.1. Simulation Setup

During Matlab simulation, it was assumed that incoming far-end speech signals are segmented into 64 ms frames with a 50% overlap between the adjacent frames. The effect of background noise and local noise has been neglected. Room impulse response (RIR) is generated using the mirror image method with the room size of $10.4 \times 10.4 \times 4.2$ m. The reflection coefficient of the wall is selected as 0.8. Room parameters, such as volume, absorption, reflections from the walls, construction materials present in the room and distance between sources and the receiver, are all taken into consideration while generating RIR. Between the clean speech signal and RIR, convolution is done by considering the fixed distance between the source and the microphones. For audio recorded by several speakers, sampling is performed at 25 kHz. The experiment is conducted using audio data obtained from the GRID corpus database. Table 1 shows all choices of sentences from GRID [33]. The corpus consists of high-quality audio recordings of around 1000 sentences spoken by each of 34 talkers (18 males, 16 females).

Table 1
Possible choices in the sentences

Commands	Colors	Prepositions	Letters	Numbers	Adverbs
Bin (b)	Blue (b)	At (a)	A–Z	1–9 and	Again (a)
Lay (l)	Green (g)	By (b)	excluding W	zero (z)	Now (n)
Place (p)	Red (r)	In (i)			Please (p)
Set (s)	White (w)	With (w)			Soon (s)

Results are obtained by calculating echo to near-end signal ratio (ENR), which is the power ratio between the echo signal and the near-end signal. It is measured by varying the distance between the microphone and the source by keeping the source fixed and moving the microphone to various positions to generate different ENR values.

5.2. Performance Evaluation

This section gives experimental results that are conducted at different ENRs: –1, –3, – and –5 dB. The Mean Opinion Score (MOS) [34] obtained by evaluating the quality of speech after EC using the proposed NMF method, and its comparison with LMS, NLMS and RLS, are plotted in Fig. 5. As ENR decreases from –1 to –5 dB, MOS increases. Also when decreasing the ENR for all methods mentioned, MOS obtained using the NMF method proposed has the highest value.

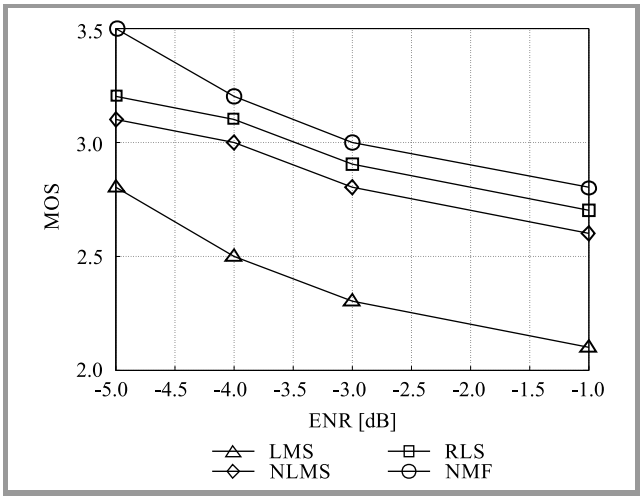

Fig. 5. MOS at different ENRs for LMS, NLMS, RLS and NMF.

5.3. Perceptual Evaluation of Speech Quality

Speech quality is evaluated using perceptual evaluation of speech quality (PESQ), and results with different ENRs are shown in Fig. 6. In this method the test signal and the corresponding reference signals are taken as input, and a set of features is extracted from both signals. They are then compared in perceptual space [35] by time-aligning these signals. Then, the speech signal is analyzed on a sample-by-sample basis for both the reconstructed output signal and the reference signal, after time-aligning it individually. This is done to compensate for any time shifts that can occur

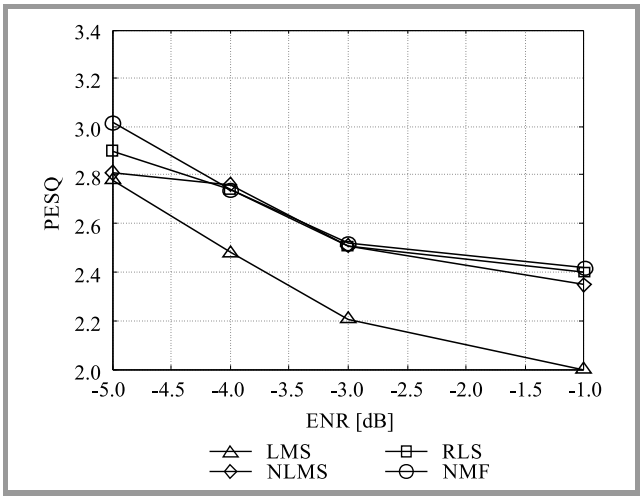

Fig. 6. PESQ at different ENRs for LMS, NLMS, RLS and NMF.

during processing. Then, the perceptual model incorporates a time frequency analysis procedure.

The experiment performed using test and reference signals is divided into 32 ms frames. The overlapping rate of the successive frames is 50%. Signals are transformed into the frequency domain using STFT. Linear frequency scale is transformed to the Bark scale for finer frequency resolution at lower rather than higher frequencies. From the perceived audio, audible differences in both domains are subtracted and accumulated over time. These are then weighted based on whether the distortion is additive in nature or whether the signal is missing. PESQ describes the audio quality using the scale of 1 (bad) to 5 (excellent). A higher PESQ value shows that the algorithm used is better suited to conduct echo cancelation [35]. From Fig. 6 it can be seen that NMF has a PESQ value that is approximately equal to or higher than all other methods, indicating a better perceptual similarity between clean and reconstructed signals. As ENR decreases from –1 to –5 dB, PESQ increases for the individual methods. Also, comparing the NMF method proposed for EC with LMS, NLMS and RLS, the PESQ value is the highest using NMF. Performance measured at ENR –5 dB obtained using NMF is found to be better than at ENR –1 dB.

5.4. Echo Return Loss Enhancement

Echo return loss enhancement (ERLE) measures the amount of additional signal loss applied by the echo canceler. It is defined as the ratio between send power in one direction and the power of a residual error signal obtained after the echo is canceled, i.e. in the steady state. It is usually measured in decibels:

$$\text{ERLE} = 10\log\frac{E\left(x^2(t)\right)}{E\left(e^2(t)\right)}, \tag{13}$$

where $x(t)$ is the send signal and $e(t) = \hat{u}(t) - u(t)$ is the residual error signal obtained after processing.

ERLE depends on the size of the adaptive filter and the algorithm which is designed to remove the echo. ERLE provides information about the behavior of the echo canceler convergence factor. The echo canceler system's input signal is an audio signal which is non-stationary in nature. In the case of adaptive filters, it is difficult to change the step size at a fast rate, which makes its implementation difficult. The NMF algorithm-based EC method presented herein overcomes those shortcomings. Figure 7 shows plots of ERLE at different ENRs. Results obtained for ERLE using NMF are higher than in the case of LMS, NLMS and RLS. As ENR decreases from –1 to –5 dB, the echo in the signal also decreases, thus the result obtained for ERLE increases. At a given ENR, statistical NMF-based EC methods produce better results. A higher ERLE indicates that speech if affected by lower echo rates. For example, at –3 dB, LMS produces ERLE 5, while NLMS 6 and RLS 7.2. The proposed NMF approach offers the result of 8.9, which is very high compared to other values. This clearly indicates that NMF has the least echo and, hence, clear audio is achieved.

5.5. Log Spectral Distortion

Log spectral distortion (LSD), also known as log spectral distance, measures the distance, in decibels, between the two spectra. It is proved in [35] that it is well suited for evaluation of algorithms. It is determined by the RMS value of the difference between log spectra of the original clean audio signal $x(n)$ and the signal that has to be evaluated $y(n)$, which is the outcome of the processing of the signal.

Let $\mathbf{S}(\omega)$ and $\hat{\mathbf{S}}(\omega)$ represent two power spectra. Then, $d(\mathbf{S}, \hat{\mathbf{S}}\omega)$ is the distance measure. LSD is given by:

$$d(\mathbf{S},\hat{\mathbf{S}})^P = (d_p)^P = \int_{-\pi}^{\pi} \left[\log\mathbf{S}(\omega) - \log\hat{\mathbf{S}}(\omega)\right]^P \frac{\mathrm{d}\omega}{2\pi}. \tag{14}$$

For $p = 1$, the above equation defines the mean absolute LSD and for $p = 2$, defines the LSD root mean square,

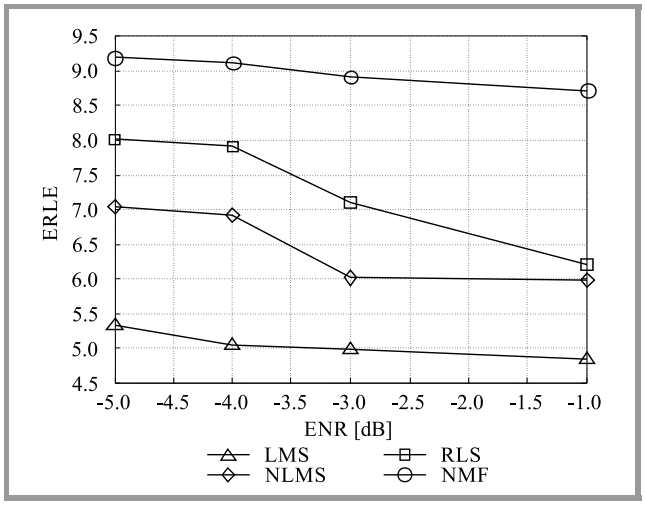

Fig. 7. ERLE vs. ENR for LMS, NLMS, RLS and NMF.

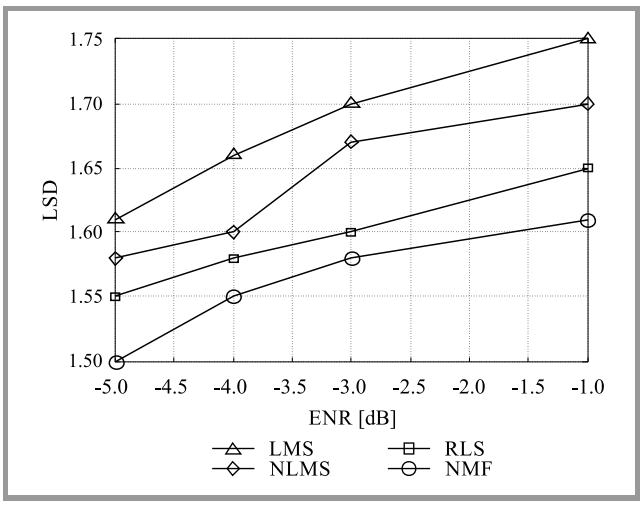

Fig. 8. LSD vs. ENR for LMS, NLMS, RLS and NMF.

which has applications in many speech processing systems and finally at $p = \infty$, the equation reduces to peak LSD. The lower the LSD value, the better the performance. Figure 8 shows LSD plots at different ENRs. At a given ENR, LSD results obtained for NMF are lower than for LMS and NLMS algorithms. At ENR –5 dB, there is less echo and thus the results obtained are better than at ENR –1 dB.

5.6. Comparison of Experimental Results

Table 3 illustrates comparisons between the proposed NMF method used for echo cancelation and LMS, NLMS and RLS algorithms using MOS, LSD, ELRE and PESQ at different ENRs. One may notice that NMF provides better EC and near-end speaker extraction. The proposed algorithm utilizes both restricted and non-restricted NMF, which results in a clean audio signal.

Table 2
Comparison of experimental results

		LMS	NLMS	RLS	NMF
ENR = –1 dB	MOS	2.1	2.6	2.7	2.8
	LSD	1.75	1.70	1.65	1.61
	ELRE	4.83	5.98	6.2	8.72
	PESQ	2	2.35	2.40	2.42
ENR = –3 dB	MOS	2.3	2.8	2.9	3.0
	LSD	1.70	1.67	1.60	1.58
	ELRE	4.98	6.01	7.10	8.90
	PESQ	2.21	2.51	2.51	2.52
ENR = –4 dB	MOS	2.5	3.0	3.1	3.2
	LSD	1.66	1.60	1.58	1.55
	ELRE	5.03	6.92	7.9	9.11
	PESQ	2.48	2.76	2.74	2.74
ENR = –5 dB	MOS	2.8	3.1	3.2	3.5
	LSD	1.61	1.58	1.55	1.50
	ELRE	5.31	7.03	8.01	9.18
	PESQ	2.78	2.81	2.9	3.02

6. Conclusion

This paper proposes a method of AEC and near-end speaker extraction using statistical NMF methods and compares it against classic LMS, NLMS and RLS methods. The results obtained with the use of the proposed NMF technique for AEC indicate that it can be implemented in real-time scenarios. Moreover, the proposed method paves the way for it to be implemented in real time scenarios using multi-resolution NMF, by taking into consideration spatial cues and reverberations. It is expected to achieve better results and to offer more accurate EC.

References

[1] E. Hansler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. John Wiley & Sons, 2005 (ISBN: 978-0-471-45346-8).

[2] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, 1st ed. Springer, 2010 (doi: 10.1007/978-1-84996-056-4).

[3] M. M. Sondhi and D. A. Berkley, "Silencing echoes on the telephone network", *Proc. of the IEEE*, vol. 68, no. 8, pp. 948–963, 1980 (doi: 10.1109/PROC.1980.11774).

[4] C. Paleologu, J. Benesty, and S. Ciochină, "An improved proportionate NLMS algorithm based on the L0 norm", in *Proc. IEEE Int. Conf. on Acoust., Speech and Sig. Process. ICASSP 2010*, Dallas, TX, USA, 2010, pp. 309–312 (doi: 10.1109/ICASSP.2010.5495903).

[5] S. Malik and G. Enzner "Recursive Bayesian control of multichannel acoustic echo cancellation", *IEEE Sig. Process. Lett.*, vol. 18, no. 11, pp. 619–622, 2011 (doi: 10.1109/LSP.2011.2166385).

[6] S. L. Gay and J. E. Benesty, Eds., *Acoustic Signal Processing for Telecommunication*. Boston: Kluwer Academic, 2000 (doi: 10.1007/978-1-4419-8644-3).

[7] J. Gunther and T. Moon "Blind acoustic echo cancellation without double-talk detection", *IEEE Worksh. on Appl. of Sig. Process. to Audio and Acoust.*, New Paltz, NY, USA, 2015 (doi: 10.1109/WASPAA.2015.7336925).

[8] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, Eds., *Advances in Network and Acoustic Echo Cancellation*. New York: Springer-Verlag, 2001 (ISBN: 978-3-662-04437-7).

[9] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement", *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008 (doi: 10.1109/TASL.2007.911054).

[10] A. H. Abdullah, M. I. Yusof, and S. R. M. Baki, "Adaptive noise cancellation: a practical study of the least-mean square (LMS) over recursive least-square (RLS) algorithm", in *Proc. Student Conf. on Res. and Develop. SCOReD 2002*, pp. 448–452, Shah Alam, Malaysia, 2002 (doi: 10.1109/SCORED.2002.1033154).

[11] X. Wang, T. Shen, and W. Wang, "An approach for echo cancellation system based on improved NLMS algorithm", in *Proc. Int. Conf. on Wirel. Commun., Netw. and Mob. Comput. WiCom 2007*, Shanghai, China, 2007 (doi: 10.1109/WICOM.2007.708).

[12] S. Wu, X. Qiu, and M. Wu, "Stereo acoustic echo cancellation employing frequency-domain preprocessing and adaptive filter", *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 3, pp. 614–623, 2011 (doi: 10.1109/TASL.2010.2052804).

[13] M. H. Maruo, J. C. M. Bermudez, and L. S. Resende, "Statistical analysis of a jointly optimized beamformer-assisted acoustic echo canceller", *IEEE Trans. on Sig. Process.*, vol. 62, no. 1, pp. 252–265, 2014 (doi: 10.1109/TSP.2013.2284138).

[14] R. C. Nongpiur and D. J. Shpak, "Maximizing the signal-to-alias ratio in non-uniform filter banks for acoustic echo cancellation", *IEEE Trans. on Circuits and Syst.: Regular Papers*, vol. 59, no. 10, pp. 2315–2325, 2012 (doi: 10.1109/TCSI.2012.2185333).

[15] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications.* Wiley-IEEE Press, 2006 (ISBN: 978-0-471-74109-1).

[16] T. S. Wadaand and B.-H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation", *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, no. 1, pp. 175–189, 2012 (doi: 10.1109/TASL.2011.2159592).

[17] S. K. Nagendra and V. S. Kumar, "Echo cancellation in audio signal using LMS algorithm", in *Nat. Conf. on Recent Trends in Engin. and Technol. NCRTET 2011*, Anand, Gujarat, India, 2011.

[18] U. Mahbub, S. A. Fattah, W-P. Zhu, and M. O. Ahmad, "Single-channel acoustic echo cancellation in noise based on gradient-based adaptive filtering", *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 20, pp. 1–16, 2014 (doi: 10.1186/1687-4722-2014-20).

[19] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values", *Environmetrics*, vol. 5, pp. 111–126, 1994 (doi: 10.1002/env.3170050203).

[20] A. Bansal, S. Choukse, K. Nathwani, and R. M. Hegde, "Acoustic echo cancellation using a multi-resolution non-negative matrix factorization method", in *Proc. 22nd Nat. Conf. on Commun. NCC 2016*, Guwahati, India, 2016, pp. 1–5, 2016 (doi: 10.1109/NCC.2016.7561119).

[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", in *Advances in Neural Information Processing Systems 13. Proceedings of the 2000 Conference*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[22] P. O'Grady, "Sparse Separation of Underdetermined Speech Mixtures", Ph.D. Dissertation, Hamilton Institute, National University of Ireland Maynooth, Ireland, 2007 [Online]. Available: https://www.hamilton.ie/publications/ogrady2007_phd.pdf

[23] S. Ciochină, C. Paleologu, J. Benesty, and C. Anghel, "An optimized affine projection algorithm for acoustic echo cancellation", in *Proc. Int. Conf. on Speech Technol. and Human-Comp. Dialogue SpeD 2015*, Bucharest, Romania, 2015, pp. 1–6 (doi: 10.1109/SPED.2015.7343092).

[24] C. Fevotte, 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, 2009 (doi: 10.1162/neco.2008.04-08-771).

[25] S. Roweis, "One microphone sound source separation", in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 793–799.

[26] F. Yang, M. Wu, and J. Yang,"Stereophonic acoustic echo suppression based on Wiener filter in the short-time Fourier transform domain", *IEEE Sig. Process. Lett.*, vol. 19, no. 4, pp. 227–230, 2012 (doi: 10.1109/LSP.2012.2187446).

[27] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech", in *Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process. ICASSP 2002*, vol. 1, pp. I-529–I-532, 2002, Orlando, FL, USA, 2001 (doi: 10.1109/ICASSP.2002.5743771).

[28] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. on Sig. Process.*, vol. 52, no. 7, pp. 1830–1847, 2004 (doi: 10.1109/TSP.2004.828896).

[29] C. Avendano, "Acoustic echo suppression in the STFT domain", in *Proc. IEEE Worksh. on Appl. of Sig. Process. to Audio and Acoust.*, New Paltz, NY, USA, pp. 175–178, 2001 (doi: 10.1109/ASPAA.2001.969571).

[30] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space", *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 5, pp. 1048–1062, 2005 (doi: 10.1109/TSA.2005.852012).

[31] E. A. P. Habets, S. Gannot, and I. Cohen, "Robust early echo cancellation and late echo suppression in the STFT domain", in *Proc. of 11th Int. Worksh. on Acoust. Echo and Noise Control IWAENC 2008*, Seattle, VA, USA, 2008.

[32] P. Yun-Sik and C. Joon-Hyuk, "Frequency domain acoustic echo suppression based on soft decision", *IEEE Sig. Process. Lett.*, vol. 16, no. 1, pp. 53–56, 2009 (doi: 10.1109/LSP.2008.2008571).

[33] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition", *The J. of the Acoustical Soc. of America*, vol. 120, no. 5, pp. 2421–2424, 2006 (doi: 10.1121/1.2229005).

[34] Y.-X. Wang and Y-J. Zhang, "Nonnegative matrix factorization: A comprehensive review", *IEEE Trans. on Knowl. and Data Engin.*, vol. 25, no. 6, pp. 1336–1353, 2013 (doi: 10.1109/TKDE.2012.51).

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs", in *Proc. IEEE Int. Conf. on Acoust., Speech and Sig. Process. ICASSP 2001*, Salt Lake City, UT, USA, 2001, vol. 2, pp. 749–752 (doi: 10.1109/ICASSP.2001.941023).

Pallavi Agrawal received her B.Eng. (Hons) in Electronics and Communication Engineering from RGPV, Bhopal, India and M.Tech. in Digital Communications from Maulana Azad National Institute of Technology (MANIT), Bhopal, India. She is now a Ph.D. candidate at the Electronics and Communication Department at MANIT. Her areas of interest include digital speech signal processing, digital communication and statistical signal processing.
E-mail: pallaviagrawal4@gmail.com
Department of Electronics and Communication Engineering
Maulana Azad National Institute of Technology
Bhopal (M. P.), India

Madhu Shandilya is currently working as a Professor at the Department of Electronics, MANIT, Bhopal, Deemed University and at the National Institute of Technology of India. She has more than 20 years teaching and research experience. Her areas of interest include image processing, information security and watermarking. She has more than 30 papers published in various national and international journals and in various IEEE-sponsored national and international conferences. She is a member of various Societies, such as IETE, CSI, and ISTE.
E-mail: madhu_shandilya@yahoo.in
Department of Electronics and Communication Engineering
Maulana Azad National Institute of Technology
Bhopal (M. P.), India