On Enhancing Model-Based Expectation Maximization Source Separation in Dynamic Reverberant Conditions Using Automatic Clifton Effect

Sania Gul[1], Muhammad Salman Khan[2], Syed Waqar Shah[1], Jaime Lloret[3,*, †]

[1]Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan.

[2]Department of Electrical Engineering, Jalozai Campus, University of Engineering and Technology, Peshawar, Pakistan.

[3]Instituto de Investigación para la Gestión Integrada de Zonas Costeras, Universitat Politècnica de València, Spain.

Email: saniagul@hotmail.com, salmankhan@uetpeshawar.edu.pk, waqar.shah@uetpeshawar.edu.pk, jlloret@dcom.upv.es

Abstract: *Source separation algorithms based on spatial cues generally face two major problems. The first one is their general performance degradation in reverberant environments and the second is their inability to differentiate closely located sources due to similarity of their spatial cues. The latter problem gets amplified in highly reverberant environments as reverberations have a distorting effect on spatial cues. In this paper, we have proposed a separation algorithm, in which inside an enclosure, the distortions due to reverberations in a spatial cue based source separation algorithm namely model-based expectation-maximization source separation and localization (MESSL) are minimized by using the Precedence effect. The Precedence effect acts as a gatekeeper which restricts the reverberations entering the separation system resulting in its improved separation performance. And this effect is automatically transformed into the Clifton effect to deal with the dynamic acoustic conditions. Our proposed algorithm has shown improved performance over MESSL in all kinds of reverberant conditions including closely located sources. On average, 22.55% improvement in SDR (signal to distortion ratio) and 15% in PESQ (perceptual evaluation of speech quality) is observed by using the Clifton effect to tackle dynamic reverberant conditions.*

Keywords: blind source separation; reverberation; dynamic acoustic conditions; Precedence effect; Clifton effect

---

*Correpondance to: Jaime Lloret, Universitat Politècnica de València, Spain.

†Email: jlloret@dcom.upv.es

I. Introduction

Source separation systems have been in use for long in many areas e.g. image processing, entertainment, medical, defence and communication systems to separate the signal of interest from unwanted signals [1]. In the field of communication, the use of source separation is mainly focused on speech enhancement (e.g. in hands free communication, Voice over IP, hearing aids, local and long distance communication [2]) and speech recognition (e.g. digital assistants, voice signature for banking transactions on call [3], security system etc). While machines are far from being excellent, human beings have an unmatched capability of sound source separation. They can easily separate sounds of their interest in the presence of other competing sounds (cocktail party problem (CPP) [4]). They desire to impart this capability to machines and the motivation behind it is to design better hearing aids, to design robust automatic speech recognition (ASR) systems, to transcribe old folk music and minutes of meeting from the available audio recordings, to improve quality of communication systems and multiple other everyday applications.

There are two approaches generally adopted for acoustic source separation. The first approach uses psychoacoustic techniques and the second approach uses purely mathematical techniques. These psychoacoustic techniques are grouped under the label computational auditory scene analysis (CASA) [5]. CASA refers to computational algorithms used for separating a sound of interest from the mixture by using psychophysical and perceptual mechanisms used by human beings to solve the CPP. CASA based approaches require just one or two mixtures to separate sources like human beings who require one or two mixtures collected at their ears to do the same task. However, listening with two ears has shown to improve intelligibility by approximately 5% as compared to listening with one ear in all types of acoustic conditions [6].

Human brain separates the sound sources by estimating their direction of arrival (DOA). The same technique is used in communication systems for the separating unknown wide band sources encountered in astronomy and unauthorized transmissions [7] and in smart antennas for enhancing the desired signal quality and reducing co channel interference [8]. Although there are many CASA based approaches utilizing the DOA for source separation, here we will focus on one of them i.e. MESSL.

MESSL [9] is a probabilistic time-frequency (TF) masking source separation system utilizing only two mixtures to separate many sources. Like human beings, in MESSL, the DOA of active sources is estimated from the two spatial cues, the i.e. interaural level difference (ILD) and the interaural time difference (ITD) (estimated from interaural phase difference (IPD)). MESSL uses the expectation maximization (EM) algorithm to cluster the spatial cues extracted from the mixture into the individual sources present inside the room. The EM algorithm is a two-step iterative process in which the process switches back and forth between expectation and maximization steps. After a number of iterations, best fit (maximum likelihood) parameters are estimated for each cluster. MESSL deals with the problem of reverberations present inside the mixture signal by introducing the concept of ‘Garbage Source’ (GS) in its model. Such a garbage source is a virtual source. All TF units that are not nearer to the mean values of ILD and ITD of any real source are allocated to the GS resulting in quality improvement of all the real sources present inside the room. However, there is a general trend of performance degradation in MESSL due to reverberations especially at small separation angles between the sources, due to distortion of spatial cues by such conditions.

In the past, many attempts have been made to improve the performance of MESSL. In [10], the authors proposed Student's t-distribution for IPD and ILD of each source resulting in improved performance over the original model using Gaussian distribution for these spatial cues. The heavy tailed t-distribution covers the outliers of non-stationary speech much better than the Gaussian distribution. In [11], the authors replaced the maximum likelihood estimation (MLE) clustering algorithm with variational Bayesian (VB) clustering, overcoming the problems of singularity and over-fitting associated with MLE algorithm and improving separation especially when sources are in close proximity. In [12] and [13], the authors used video in addition to audio mixtures to improve the source localization capability of MESSL while in [14]; they implemented spatial covariance in the EM algorithm in their previously proposed model of [13]. In [15] and [16], the authors in addition with audio and video, used a circular beamformer instead of the two microphone design of [12] to reduce the uncertainties in source localization. With the recent advancement in artificial neural networks (ANNs), many researchers have attempted to perform source separation using neural networks and found the results quite pleasing. In [17] and [18], beamforming and deep neural networks (DNN) are used for improving performance of MESSL. These systems use both the spectral and spatial cues for source separation. However these systems need more than two mixtures and a lot of training before they can successfully perform the source separation task.

In order to improve the performance of MESSL in a dynamic reverberant environment, in this paper we propose a method which restricts the reverberations entering MESSL by using the Clifton effect. The Clifton effect is defined as the dynamic component of the precedence effect which adapts the precedence effect to the changing acoustic conditions. The precedence effect [19] is an auditory mechanism that aids humans in localizing sounds in reverberant environments by giving more perceptual weight to the direct sound compared to the later reflections (reverberations). These late reflections distort the location information (spatial cues) present inside the sound and thus make it difficult for the listener to localize the source. This is not only true for sound signals, but for any indoor wireless communication, where non line of sight (NLOS) signals (reflections) can easily produce errors in location tracking [20]. It is found that, if the acoustic conditions are changed, the precedence effect adapts itself according to the new conditions in which the listener is situated. This adaptation is caused by the dynamic component of precedence effect, called as the Clifton effect.

Our model is inspired from the work of [21] which models the Clifton effect in its source separation system. However, their model parameters do not adapt themselves automatically according to the changing acoustic conditions, so we can call it Pseudo-Clifton effect model. In [21], the speech band (50 Hz to 8000 Hz) is divided into a fixed number of sub bands. Each sub-band is called a 'channel' or a 'frequency strand'. These channels are subjected to low pass filtering to suppress the reverberations. In [21], the number of frequency strands or channels over which low pass filtering is applied to filter out the rapid fluctuations caused by reverberations are fixed and the parameters of the low pass filter are adapted to match the changing acoustic conditions. Contrarily in our proposed model, the number of frequency strands over which low pass filtering is applied, is changed every time with the changing acoustic conditions while the parameters of the low pass filter are kept fixed. In [21], the interaural cues of individual sources are grouped on the basis of their common azimuth, while in our proposed model this grouping is done by using the EM algorithm. As already mentioned above the model of [21] cannot update its parameters automatically according to the changing acoustic conditions and needs manual up gradation of parameters when the acoustic conditions are changed. But the model parameters are adjusted automatically according to the changes in acoustic conditions in our proposed algorithm.

The rest of the paper is organized as follows. In the following section, we will focus on the Clifton effect and its usefulness in dynamic acoustic conditions. We will give our proposed system overview in section III. Our algorithm is summarized in section IV. In section V, we will describe the experimental setup, the evaluation criteria and the comparison methods. We will present experimental results and comparison statistics of different models in section VI. We will compare our proposed algorithm with other algorithms in section VII and conclude the paper in section VIII.

II. Related Work

It is observed that source separation performance is improved if the speech mixtures are dereverberated before separation. A lot of research has been done so far in this regard. For example, the authors in [22] improved their source separation model proposed in [12], by dereverberating the audio mixtures by subtracting spectral contents of late reflections and in [23] by subtracting in cascade the spectral and the power contents of late reverberations from the speech mixtures. As our proposed model utilizes the Clifton effect, so we will put more focus here on those models which utilize this effect for deverberation of speech mixtures.

The authors in [24] proposed a model for source separation which uses the Precedence effect or the law of first wave front to tackle the issue of reverberation. Their model inhibits the reverberations, using only the initial unaffected wave fronts for source separation. However, this model had fixed 'inhibitory parameters', which restricts its usefulness in all kinds of acoustic conditions.

So, in [21], the authors pointed out the need to change the parameters of the precedence effect model proposed in [24] with the changing acoustic conditions. Different rooms have different acoustic conditions e.g. different ITDGs (Initial Time Delay Gaps), DRRs (Direct to Reverberant Ratios) and RT_{60}s (Reverberation Times). The authors proposed to use room-specific components (the inhibitory parameters: the inhibitory time constant (α_p) and the inhibitory gain (G)) in the precedence effect model which would improve separation performance. Inhibitory time constant decides how early the inhibition must start after the onset and the inhibitory gain decides about the strength of inhibition. In the Precedence effect model proposed in [24], the default values of these parameters are G=1 and α_p=15ms**.** In [21], the authors carried out experiments in five different rooms which they labeled "X", "A", "B", "C" and "D", all having different acoustic conditions and found the most optimum inhibitory parameters for each room (details in [21]).

The authors have not implemented the Clifton effect in its true sense, as their model parameters do not adapt to the changing acoustic conditions automatically as discussed in [21]. This restricts the model's usefulness in all those applications which involve user mobility from one acoustic condition to another. The system performance degrades in such situations as the model parameters are not updated to the new conditions automatically and require manual settings according to the new conditions. In all such applications, a self-adapting model is required which can blindly extract the acoustic parameters of the room without any prior information and adjust its system parameters accordingly. Unfortunately, few datasets are available representing the neurophysiological mechanisms that are responsible for the automatic adaptation of the dynamic component of the precedence effect according to the acoustic conditions, although the authors in [25] agree that this effect is partially achieved via feedback from the higher auditory systems to peripheral auditory systems through the centrifugal pathways. We will

therefore utilize this feedback concept in our proposed model to automatically adjust the parameters according to the dynamic acoustic conditions.

III. System Overview

Our proposed system consists of two main blocks connected in cascade with each other as shown in Figure 1. The front end consists of the de-reverberation block, which blocks the incoming echoes in dynamic reverberant conditions by utilizing the Clifton effect. The backend consists of the source separation block, which performs the speech separation task on the two mixtures that it receives from the de-reverberation block. We will term the proposed model CMESSL. In this acronym 'C' represents the Clifton effect and 'MESSL' represents model-based expectation-maximization source separation and localization algorithm from which the source separation block of our proposed model is inspired.

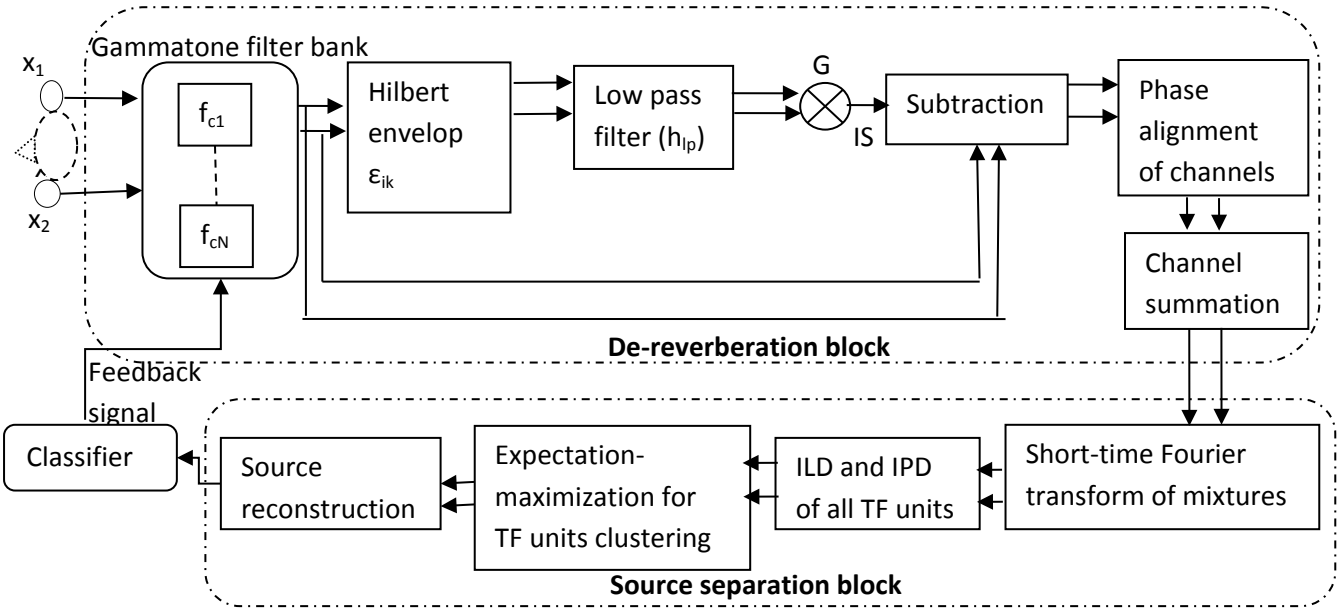

Figure 1: Block diagram of CMESSL, showing the application of the Clifton-effect on speech mixtures followed by source separation process

Assume that the number of active sources W present inside the room is known a priori and there are k microphones, recording the audio mixtures where k = {1, 2}. Each source s_i is convolved with the room impulse response (RIR) h_{ki} that exists between the source s_i and the k^{th} microphone and added together with other convolved sources to form mixture x_k at the k^{th} microphone. The process of mixture formation is shown in Figure 2 below and is given at the k^{th} microphone by equation (1) as:

$$\boldsymbol{x_k}(t_s) = \sum_{i=1}^{W} s_i(t_s) * h_{ki}\,(t_s) \qquad (\mathbf{1})$$

where '*'represents the convolution operation and t_s represents the sampling time index when the sampling frequency at which mixture samples are taken is f_s. The assumption is made that except for the active *W* sources, there are no unidentified noise generators which are contributing to the mixtures.

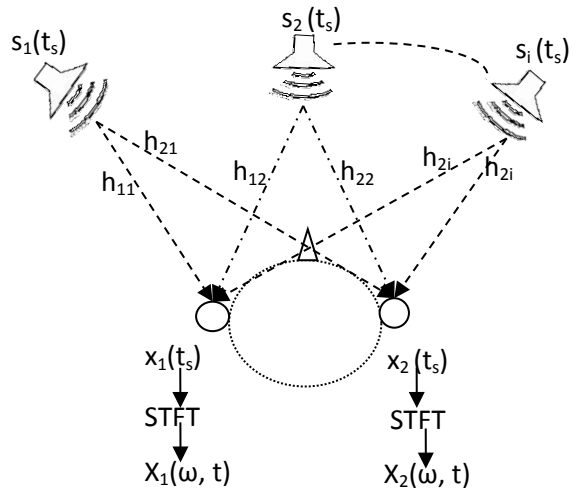

Figure 2: Signal notations. The two mixture signals are transformed to the time frequency domain $X_1(\omega, t)$ and $X_2(\omega, t)$ by STFT.

The two mixtures are then entered the de-reverberation block where each mixture passes through the Gammatone Filter Bank (GTFB). It is a fourth-order filter bank [26] with each filter having impulse response in continuous time domain given by equation (2)

$$g(t) = t^3 e^{-2\pi bt}\cos(2\pi f_0 t)u(t), t \geq 0 \quad (\mathbf{2})$$

where f_0 is the centre frequency, u(t) is the unit step function and *b* is the bandwidth parameter [21] and t represents continuous time. The purpose of the gammatone filter bank is to select the frequency strands or the channels over which low pass filtering will be applied later on. These channels are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale [27]. Number of channels (*N*) is the parameter providing Clifton-effect in our model. Each channel *i* which comes out of the GTFB is represented by $\mathcal{H}_{i,k}$ where *k* is the mixture index. After this stage, the instantaneous Hilbert envelop $\varepsilon_{ik}(n)$ of each channel is extracted as in [21] and this envelope is passed through the Low Pass Filter (LPF) with impulse response h_{lp} to create the inhibitory signal (*IS*) as given in equation (3):

$$IS(n) = h_{lp}(n) * \varepsilon_{ik}(n) \quad (\mathbf{3})$$

where $h_{lp} = \mathrm{A}ne^{-\frac{n}{\alpha_p}}$, α_p is inhibitory time constant, A is set to have unity gain at zero frequency, '*' represents the convolution operation and n is the discrete time index. This inhibitory signal is afterwards multiplied by inhibitory gain factor *G*, and later subtracted from the corresponding GTFB channel $\mathcal{H}_{i,k}$ to obtain the output channel $r_{i,k}$ as shown in equation (4):

$$r_{i,k}(n) = [\mathcal{H}_{i,k}(n) - G(IS(n))]^+ \quad (\mathbf{4})$$

The superscript "+" represents half wave rectification. The subtraction operation results in suppression of steady state portion of channel $r_{i,k}$ leaving only the signal transients which contain localization cues of the source. Finally, all channels corresponding to each mixture are phase-aligned and added together again creating two mixtures which will then enter the source separation block.

In the source separation block, the mixtures are first converted from the time domain to the time frequency (TF) domain by taking their STFT (Short Time Fourier Transform) as in (5):

$$X_k(\omega,t) = \mathcal{F}(\boldsymbol{x_k}) \qquad (\mathbf{5})$$

Then the interaural spectrogram is obtained by taking the ratio of the two mixtures $X_1(\omega, t)$ and $X_2(\omega, t)$ at each TF point as shown in equation (6) below.

$$\frac{X_1(\omega,t)}{X_2(\omega,t)} = 10^{\frac{\alpha(\omega,t)}{20}} e^{j\phi(\omega,t)} \qquad (\mathbf{6})$$

where $\alpha(\omega,t)$ is the ILD in dB and $\phi(\omega,t)$ is the observed IPD. IPD must lie in the region {-π, π} to avoid spatial aliasing. We will use $\theta^{GS}{}_{\Omega\Omega}$ mode of MESSL [9] in our source separation block, in which both ILD and IPD are modeled as frequency dependent parameters and the Garbage Source (GS) is used to deal with reverberations. As given in [9], this mode performs the best among all possible modes of MESSL. The observed IPD values from the mixtures i.e. $\angle\frac{X_1(\omega,t)}{X_2(\omega,t)}$ at each TF point do not always map to correct ITD due to spatial aliasing. So a top down approach is used for the calculation of ITD, where IPD is estimated by plugging in different values of delay (τ) in the range {-15:0.5:+15} samples. The τ which produces the closest match to the observed IPD is selected. However, it is required that the delay (τ) and the length of RIR must be smaller than the STFT frame length. Any portion of RIR above the STFT frame length would be treated as noise.

The phase residual error $\hat{\phi}$ is defined as the difference between the observed IPD and estimated IPD and given in equation (7) as:

$$\hat{\phi} = \angle\frac{X_1(\omega,t)e^{-j\omega t}}{X_2(\omega,t)} \qquad (\mathbf{7})$$

$\hat{\phi}$ lies in the interval {-π, π}. Both ILD and IPD residual are modeled as normal distributions. Let $\varrho(\omega)$ and $\eta^2(\omega)$ be the mean and variance of ILD(α) and $\xi(\omega)$ and $\sigma^2(\omega)$ be the mean and variance of IPD residual ($\hat{\phi}$) respectively. Then ILD and IPD models for each source s_i at each TF point are given in (8) and (9) as:

$$p\left(\alpha(\omega,t) \,\middle|\, \varrho_i(\omega),{\eta_i}^2(\omega)\right) = \mathcal{N}\left(\alpha(\omega,t) \,\middle|\, \varrho_i(\omega),{\eta_i}^2(\omega)\right) \qquad (\mathbf{8})$$

$$p\left(\hat{\phi}(\omega,t;\tau) \,\middle|\, \xi_{i,\tau}(\omega),{\sigma_{i,\tau}}^2(\omega)\right) = \mathcal{N}\left(\hat{\phi}(\omega,t;\tau) \,\middle|\, \xi_{i,\tau}(\omega),{\sigma_{i,\tau}}^2(\omega)\right) \qquad (\mathbf{9})$$

The subscripts with mean and variance symbols in equation (8) show that the ILD distribution parameters are dependent only on frequency ω, while those in equation (9) show that the IPD parameters are dependent on both frequency ω and delay τ. Implementing the correlation that is known to exist between ILD and ITD in the respective means of these distributions and assuming conditional independence of only their noisy part, the joint probability of ILD and IPD models is given in (10) as:

$$p(\alpha(\omega,t),\hat{\phi}(\omega,t;\tau)|\hat{\boldsymbol{\theta}}) = \mathcal{N}\left(\alpha(\omega,t)\,\middle|\,\varrho_i(\omega),{\eta_i}^2(\omega)\right).\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,\middle|\,\xi_{i,\tau}(\omega),{\sigma_{i,\tau}}^2(\omega)\right) \quad (\mathbf{10})$$

where $\hat{\boldsymbol{\theta}} = \{\varrho_i(\omega),{\eta_i}^2(\omega),\xi_{i,\tau}(\omega),{\sigma_{i,\tau}}^2(\omega),\psi_{i,\tau}\}$ represents all model parameters for source s_i. $\psi_{i,\tau}$ is the mixing weight i.e. the proportion of the total TF points of mixture belonging to source s_i at delay τ in the Gaussian Mixture model which manifest itself due to mixing of many such Gaussian distributions belonging to different combinations of sources and delays. The log likelihood, given the observation $\hat{\boldsymbol{\theta}}$ over all TF points is given in (11) as:

$$L(\hat{\boldsymbol{\theta}}) = \sum_{\omega,t} \log p\left(\phi(\omega,t),\alpha(\omega,t)\,\middle|\,\hat{\boldsymbol{\theta}}\right) \quad (11)$$

Marginalizing over all sources and all delays, the log likelihood function is given as in (12).

$$L(\hat{\boldsymbol{\theta}}) = \sum_{\omega,t} \log \sum_{i,\tau} \left[\mathcal{N}\left(\alpha(\omega,t)\,\middle|\,\varrho_i(\omega),{\eta_i}^2(\omega)\right).\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,\middle|\,\xi_{i,\tau}(\omega),{\sigma_{i,\tau}}^2(\omega)\right)\right].\psi_{i,\tau}] \quad (\mathbf{11}2)$$

The Maximum Likelihood solution is given as in (13).

$$L(\hat{\boldsymbol{\theta}}) = max_\theta \sum_{\omega,t} \log p\left(\phi(\omega,t),\alpha(\omega,t)\,\middle|\,\hat{\boldsymbol{\theta}}\right) \quad (13)$$

This solution is achieved by using the expectation maximization (EM) algorithm where the system switches back and forth between E and M step, until the maximum likelihood solution is obtained or the processor completes the pre-set number of iterations. In the E step, likelihood of each TF point belonging to source i and delay τ is given as in (14)

$$v_{i,\tau}(\omega) \propto \psi_{i,\tau}.\mathcal{N}\left(\alpha(\omega,t)\,\middle|\,\varrho_i(\omega),{\eta_i}^2(\omega)\right).\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,\middle|\,\xi_{i,\tau}(\omega),{\sigma_{i,\tau}}^2(\omega)\right) \quad (14)$$

In the M step, new model parameters $\hat{\hat{\boldsymbol{\theta}}}$ are estimated from all TF points, weighted by their corresponding likelihood probabilities (calculated in E step) by the formulae given in (19) to (23) in [9]. At the end, the probabilistic mask for each source is formed as in (15)

$$M_i(\omega,t) = \sum\nolimits_\tau v_{i,\tau} \quad (\mathbf{12})$$

The desired source is extracted by applying this mask to the mixture and then inverse STFT is performed to get the time domain signal. The volume of sound is then adjusted according to the listener's requirement.

The classifier block (outside the dotted blocks) performs the automation of the Clifton effect which is needed if a user is roaming in dynamic acoustic conditions. The details of its working are given in Section VI-C below.

IV. Algorithm Summary

Our proposed model takes two mixtures collected at the microphones and de-reverberate individual channels of each mixture (selected by the gammatone filter bank (GTFB)), align and add them together to again create two mixtures which then enter the source sparation block. Here the spatial features are extracted from the mixtures which are then used to cluster the TF units of mixtures into individual sources by employing the EM algorithm. Our proposed algorithm is summarized as given below.

Input: Speech mixtures collected at two microphones.

Output: Separated speech sources.

1. Prepare the mixtures from the sources as given in equation (1) and depicted in Figure 2.
2. De-reverberate the mixtures by following the procedure given in equation (2)-(4) over the frequency strands selected by gammatone filter bank.
3. Enter these de-reverberated mixtures in source separation block and convert them to time frequency (TF) domain by applying short time Fourier transform as given above in equation (5).
4. Use PHAT algorithm [28] for initialization of certain parameters and run the EM algorithm to achieve maximum likelihood parameters $\hat{\boldsymbol{\theta}} = \{\varrho_i(\omega), {\eta_i}^2(\omega), \xi_{i,\tau}(\omega), {\sigma_{i,\tau}}^2(\omega), \psi_{i,\tau}\}$ of each source s_i.
5. This maximum likelihood solution is then used to create time frequency masks for each source as shown in equation (14) and (15) above.
6. Apply the mask on the mixture to retrieve the desired sources.

The flow chat of our proposed algorithm is given in Figure 3 below.

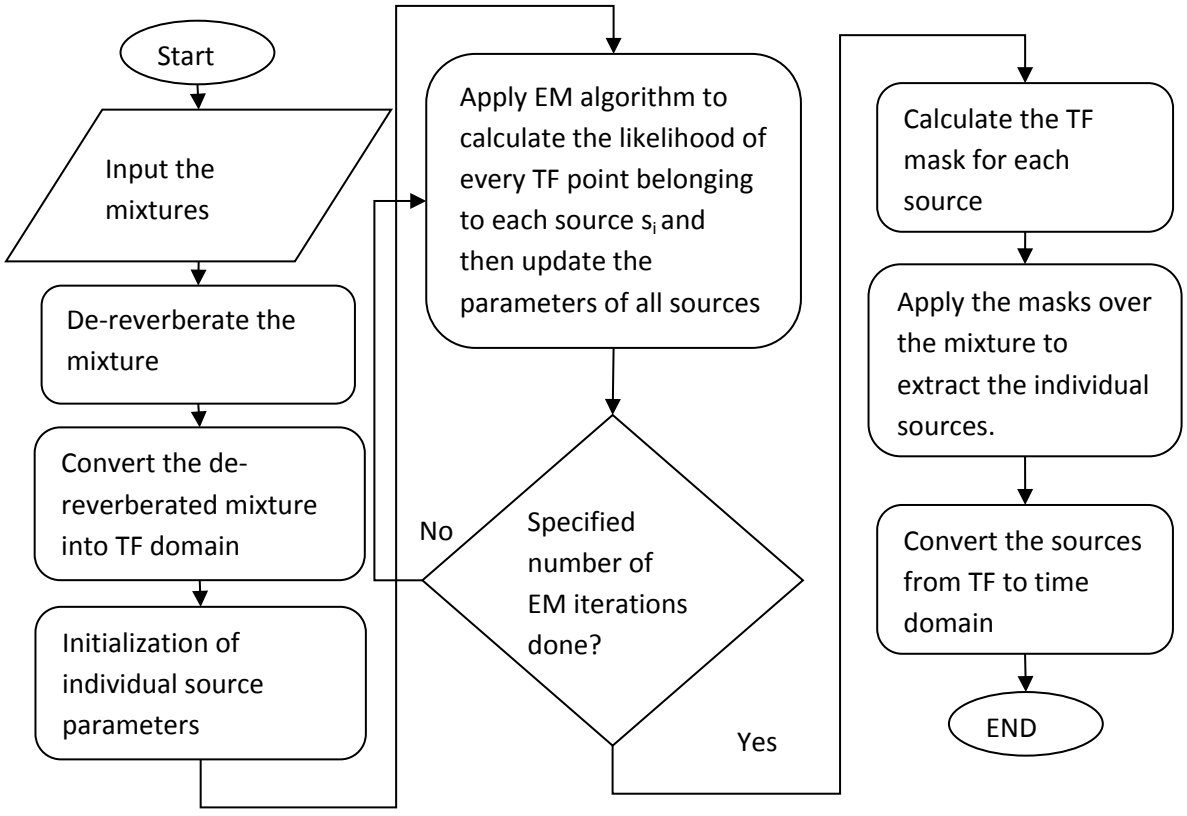

Figure 3: Flow Chart of our proposed algorithm

V. Experimental Evaluation Parameters

We will evaluate the performance of our proposed algorithm in two main sets of experiments. In the first experiment, separation performance of our proposed algorithm is compared with three other speech separation algorithms with varying separation angles. In the second experiment, the effect of applying the Clifton effect on a spatial cue based source separation algorithm at small angles is investigated. Both sets of experiments are performed under real room impulse responses. And finally, the automatic adaptation procedure, according to the acoustic conditions is discussed.

A brief detail of the experimental set-up including room layout, data set, room impulse responses (RIRs), evaluation metrics, model parameters and overview of different algorithms used in our experiments is given below.

A. Room Layout

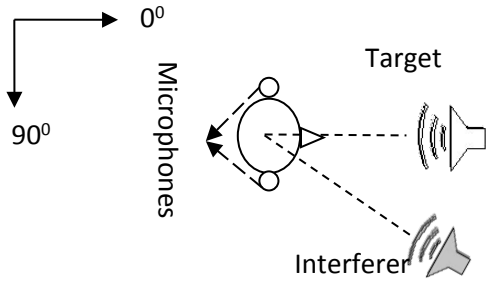

Figure 4: The room layout showing one of the approximate positions of the sources and the sensors.

We perform our experiments in six different rooms listed in Table 1. According to the room impulse responses (RIRs) used in our experiments, our equipment setup can be visualized as given in Figure 4. The distance between the two microphones is 0.14 m (equal to the average distance between two human ears) and the radial distance between the centre of microphones and each source is 1.5m in rooms {X, A, B, C, D} [21] and 1m in room {S} [29]. We perform the experiments with two sources in each room; one is target, the other interferer. So we are designing the solution for 'determined or square case' de-mixing problem where the number of simultaneous sources is equal to the number of microphones [30]. Like all other source separation systems, performance will degrade by increasing the number of sources due to decline of sparseness [31] and closeness of sources.

B. Dataset

Each speech mixture is generated by using two sources randomly selected from the TIMIT corpus [32]. This corpus contains 6300 sentences, with 10 sentences spoken by each of 630 Native American English speakers. The sources are selected carefully so that the mixtures consists of all the three possible combinations i.e. male-male, male-female and female-female. The sources are cropped to an equal length

of 2.5 s to avoid the variable silence period of each speaker at the end of the sentence. At each angular separation of target and interferer, five mixtures are separated and the results are averaged. The mixture at each microphone is generated by first convolving each source signal separately with the room impulse response (RIR) that exists between the microphone and that source, and later adding the convolved sources.

C. Real Room Impulse Responses (Real RIRs)

We will use the RIRs of the rooms given in Table 1. The$RT_{60}s$ of these rooms are in the range {0, 1} seconds, since these $RT_{60}s$ are representative of most of the real world acoustic conditions. The RIRs of rooms {X, A, B, C, D} were captured by keeping the distance of 1.5 m between the loudspeaker and HATS (Head and Torso Simulator) (details in [21]). And the RIRs of room S are recorded by keeping distance of 1m between the KEMAR (Knowles Electronic Manikin for Acoustic Research) dummy head and the loudspeaker (details in [29]). We will only use the RIRs of room S recorded in the centre of room.

D. Evaluation Criteria

We will use two objective evaluation metrics to compare our model performance with other models. These are signal-to-distortion ratio (SDR) [33] and Perceptual Evaluation of Speech Quality (PESQ) [34].

SDR measures the overall signal distortion in decibels and PESQ measures the quality of separated speech as perceived by the listener and has a value range between -0.5 to 4.5 (the higher, the better).

E. Model Parameters

Before taking readings in each room, the number of channels or frequency strands N in gammatone filter bank in the de-reverberation block of our proposed model needs to be adjusted according to the acoustic conditions. All values of N ranging from 2 to 32 are tested and the most optimum values of N providing the best separation results in each room are listed in Table 1.

Table 1: Optimum values for number of channels (*N*) in Gammatone Filter Bank in each room.

Rooms	RT_{60}(seconds)	Number of channels *(N)*
X	0	10
A	0.32	8
B	0.47	6
C	0.68	6
D	0.89	6
S	0.565	5

However, as already mentioned, there is no need to change the inhibitory parameters in each room in our proposed model and they are kept at their default values (G=1 and α_p= 15ms). The value of N is high for rooms having lower values of RT_{60}s and vice-versa. In highly reverberant situations, keeping a large value of N would cause additional distortion of spatial cues (which are already highly devastated due to echoes) by the low pass filtering to be applied on now more number of channels, resulting in increased degradation in performance. On the other hand, reducing the number of channels in filter bank below N specified for each room in Table 1 would cause insufficient cleaning of reverberations, again causing

decline in performance. The STFT parameters used in the source separation block of our proposed model are summarized in Table 2.

Table 2: STFT parameters of CMESSL

Sampling frequency	16 kHz
STFT frame length	1024 samples
Hop size	256 samples
Velocity of sound	343 m/s
Source signal duration	2.5 s

F. Overview Of Competing Algorithms

We will evaluate the performance of our proposed model CMESSL by comparing its results with three other time-frequency based source separation algorithms in [35], [21] and [9]. The model proposed in [35] is called Degenerate Un-mixing Estimation Technique (DUET). It is a binary time-frequency (TF) masking system, which assumes W-Disjoint Orthogonality of sources present inside the mixture. It can separate any number of speech sources with only two mixtures. The TF points of mixtures are clustered on the basis of attenuation and delay at each point. This system is designed for source separation in anechoic environment. Its performance degrades rapidly in presence of echoes.

The second model proposed in [21] is the Clifton source separation model which also requires only two mixtures to separate many sources. It is again a binary TF masking system which uses the Clifton effect (although not in its true sense as the system requires manual adjustment for any change in acoustic conditions) to deal with reverberations. It then performs cross-correlation of the two de-reverberated mixtures to estimate the ITD of each source. This ITD is mapped to corresponding azimuth by lookup table. ILD is calculated from the envelopes of the two de-reverberated mixtures and is used to dropout those TF units which are dominated by noise or reverberations. We will term this model CLIFTON in our future discussion.

The third method proposed in [9] is MESSL. It is a probabilistic time-frequency source separation model requiring only two mixtures to separate the sources. Its basic principle of working is already given in Section I. In this paper, we will use $\theta_{\Omega\Omega}^{GS}$ mode of MESSL, where the subscript 'ΩΩ' denotes that both ILD and ITD models of each source are frequency-dependent and the superscript 'GS' shows that garbage source is activated. We have used the term MESSL to refer to this model in our entire discussion.

VI. Experimentation and Results

A. CASE 1: Comparison with other algorithms

In this experiment, performance of four systems namely CMESSL, MESSL, CLIFTON and DUET are compared in all the rooms mentioned in Table 1. The target source is fixed at 0^o with respect to the perpendicular bisector of the axis passing through the microphones and the interferer is moving in an arc around the microphones with target-interferer angular separation starting from 15^0, incrementing in step size of 15^0 and reaching the maximum at 90^0 as depicted in Figure 4 above.

The results are given in Figures 5 to 10 below.

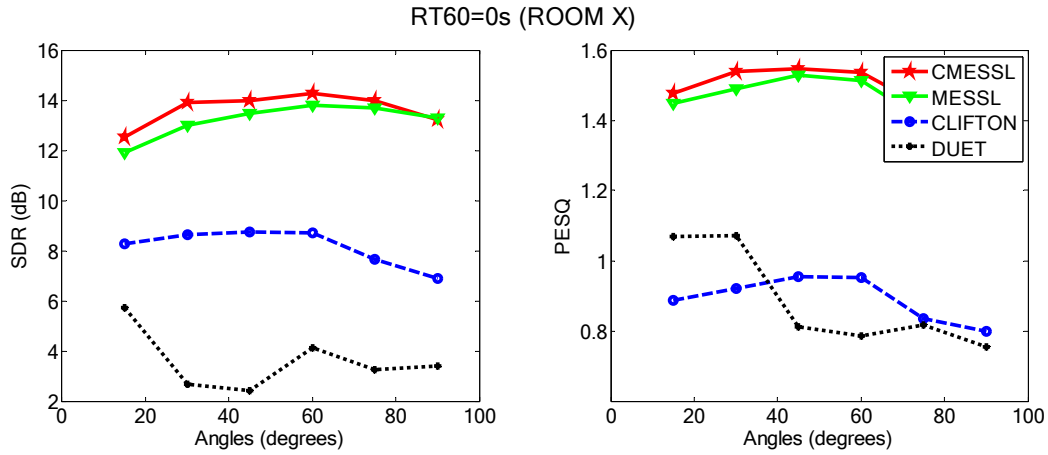

Figure 5: Comparison of algorithms in Room X with RT_{60} of 0ms.

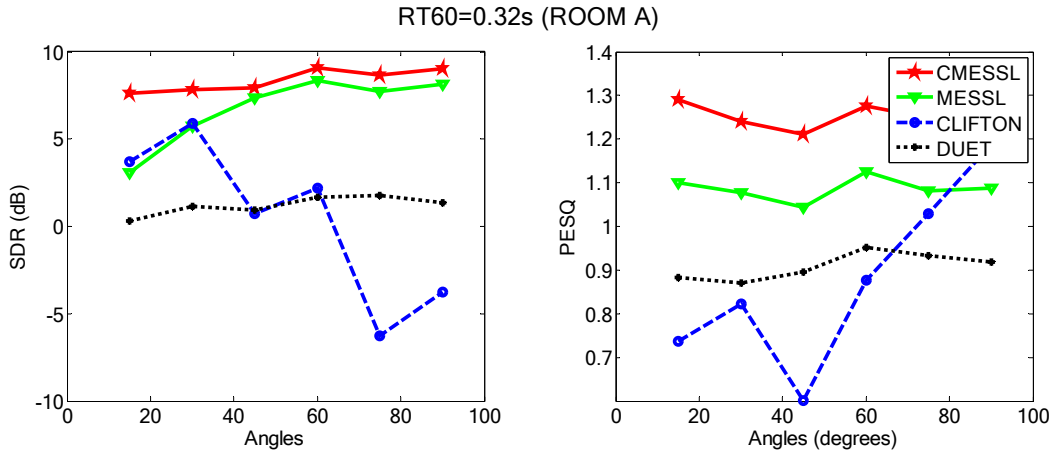

Figure 6: Comparison of algorithms in Room A with RT_{60} of 320ms.

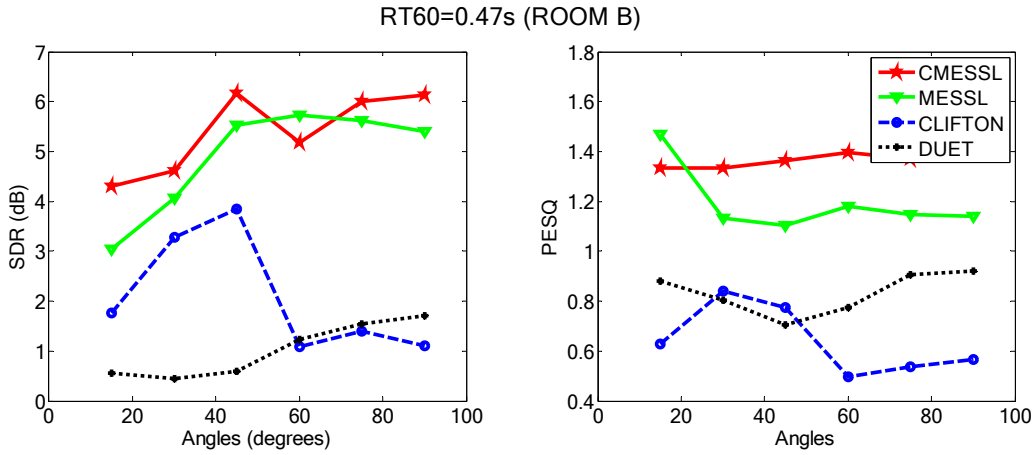

Figure 7: Comparison of algorithms in Room B with RT_{60} of 470ms.

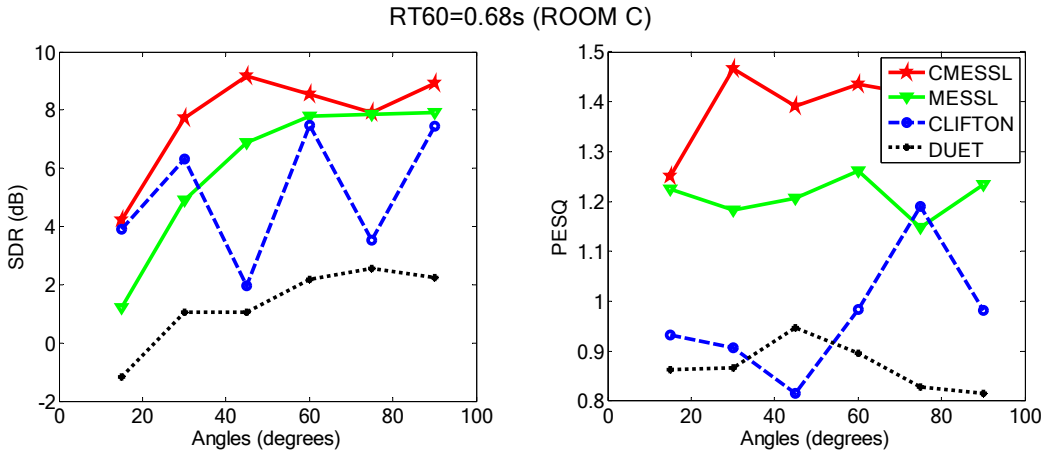

Figure 8: Comparison of algorithms in Room C with RT_{60} of 680ms.

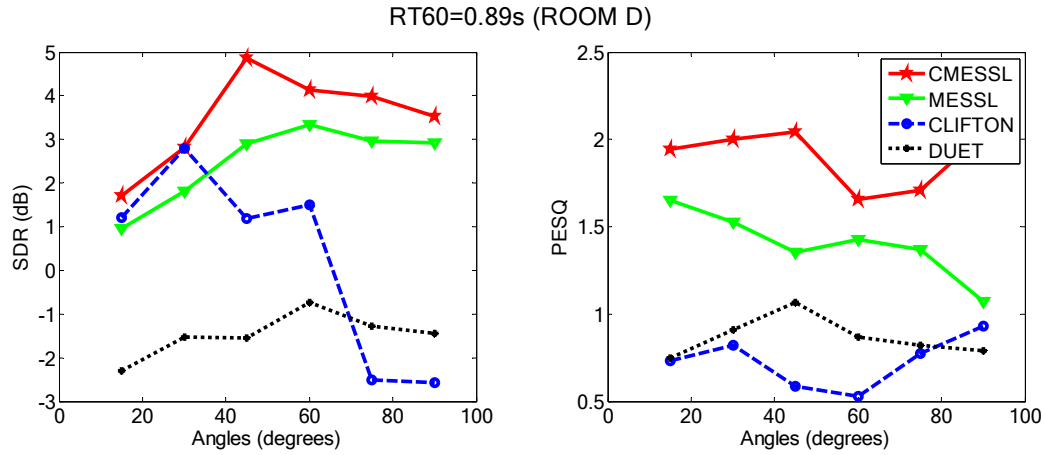

Figure 9: Comparison of algorithms in Room D with RT_{60} of 890ms.

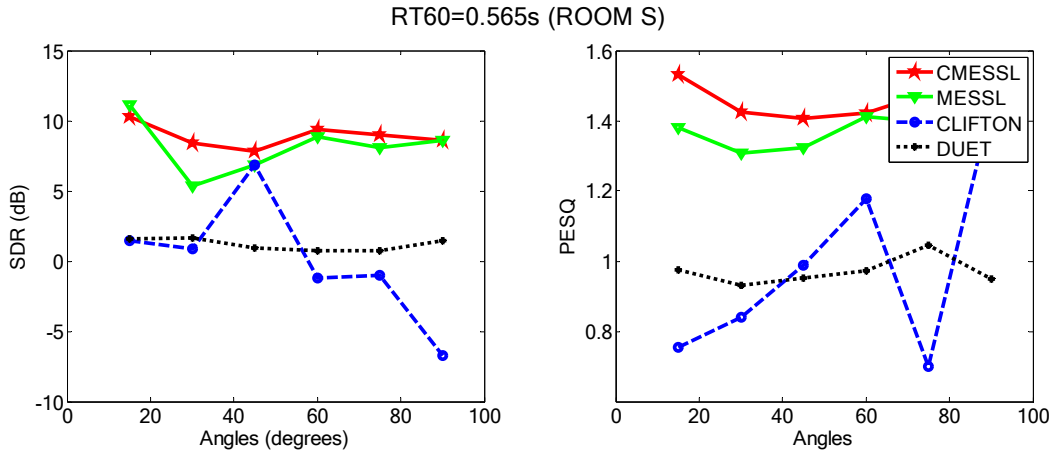

Figure 10: Comparison of algorithms in Room S with RT_{60} of 565ms.

Figure 5 shows the results of comparing our proposed algorithm CMESSL with other algorithms in dry chamber (RT_{60} = 0ms). Although our proposed model CMESSL shows improvement over other competing algorithms yet this improvement is not substantial as there are no echoes in the room which needs to be filtered out so the output is almost equal to that of MESSL. DUET performance is also good in such situation as it is designed specifically for anechoic conditions. While the performance of CLIFTON model was good at smaller separations, it declines slightly at larger separations.

Figure 6 depicts the comparative result in room A where there are low reverberations (RT_{60} = 320 ms). As there are reverberations so here our models start showing its potential by suppressing the echoes which may be detrimental to source separation performance. In such conditions, our proposed model gives an average SDR of 8.37 dB which is 1.633 dB higher than MESSL, 7.94 dB higher than CLIFTON and 7.16 dB higher than DUET. Likewise, in terms of PESQ, our proposed model has an average of 1.25, which is 0.166, 0.38 and 0.34 points higher than MESSL, CLIFTON and DUET respectively.

Figure 7 shows the results in room B where environment can be regarded as medium reverberant (RT_{60} = 470 ms). In terms of SDR, again our proposed algorithms improves SDR and PESQ at almost all separation angles, while the performance of CLIFTON shows major drop after the source-interferer separation increases beyond 45^o.

Figure 8 shows the results in room C which also comes under the medium echoic conditions (RT_{60} = 680 ms). In this room the most notable behavior is that of CLIFTON model which shows a zigzag rising trend in terms of SDR and continuous rise in terms of PESQ for increasing separations.

Figure 9 shows the comparison in room D which is the most difficult situation to handle by source separation algorithms due to presence of echoes which continue to interfere with direct path signal for much longer durations as the echo die-out time is very large (RT_{60} = 890 ms). Even in such situations, our proposed model out performs the other three algorithms both in terms of SDR and PESQ. Its average SDR in room D is 3.51 dB which is 1.02 dB, 3.24 dB and 4.98 dB higher than MESSL, CLIFTON and DUET respectively. And in terms of PESQ, our proposed model has an average of 1.8, which is 0.5, 1.17 and 1.03 points higher than MESSL, CLIFTON and DUET respectively.

Figure 10 shows the results in room S (RT_{60} = 565 ms). As clear from the figure our proposed model is better than the rest of algorithms under consideration. The average SDR of our proposed model is 8.93 dB which is 0.79 dB, 8.89 dB and 7.73 dB higher than MESSL, CLIFTON and DUET respectively and our model's average PESQ is 1.44 which is 0.08, 0.46 and 0.47 points higher than the above three mentioned models respectively.

The results above indicate that our proposed model's performance is the best in all kinds of acoustic conditions both in terms of SDR and PESQ. It gives substantial improvement over all other competing algorithms including its strongest competitor MESSL.

CLIFTON shows good performance at small separation angles but its performance drops down with increasing separation angles in all the rooms except room C. For example, in all rooms except room C, the average SDR is 3.516 dB higher at the first three smaller angles (15, 30 and 45 degrees), than the average SDR at three larger angles (45, 60 and 90 degrees). So, a general conclusion can be made that the strength of CLIFTON model is its better performance at small separation angles in almost every acoustic

condition. This is because this model uses cross correlation between the two mixtures to estimate the ITD of active sources which reduces as the separation between the sources increases resulting in performance decline at larger separation angles.

In contrast, MESSL shows better performance at larger separation angles than at the smaller ones in all the rooms as instead of cross correlation between mixtures it uses PHAT algorithm which is a more accurate source locator at separations larger than 10^o. For example, its average SDR in all six rooms at the three smaller angles i.e. 15, 30 and 45 degrees is 1.7 dB higher than the SDR at the three larger angles mentioned above. So, the strength of MESSL is its better performance at large separation angles in all types of acoustic conditions.

DUET is the worst performer among all algorithms as it faces the problems of spatial aliasing and reverberations, the conditions for which it was not designed. The SDR drops below 2 dB as soon as the RT_{60} rises above 0 ms and even slides down below 0 dB in highly reverberant conditions of room C and D.

So, we consider DUET as the baseline and measure the overall improvement achieved by other algorithms with reference to DUET. The results of improvement of all algorithms over DUET in each room are given in Figure 11 below which indicates the supremacy of CMESSL over all algorithms under consideration, in all kinds of acoustic conditions.

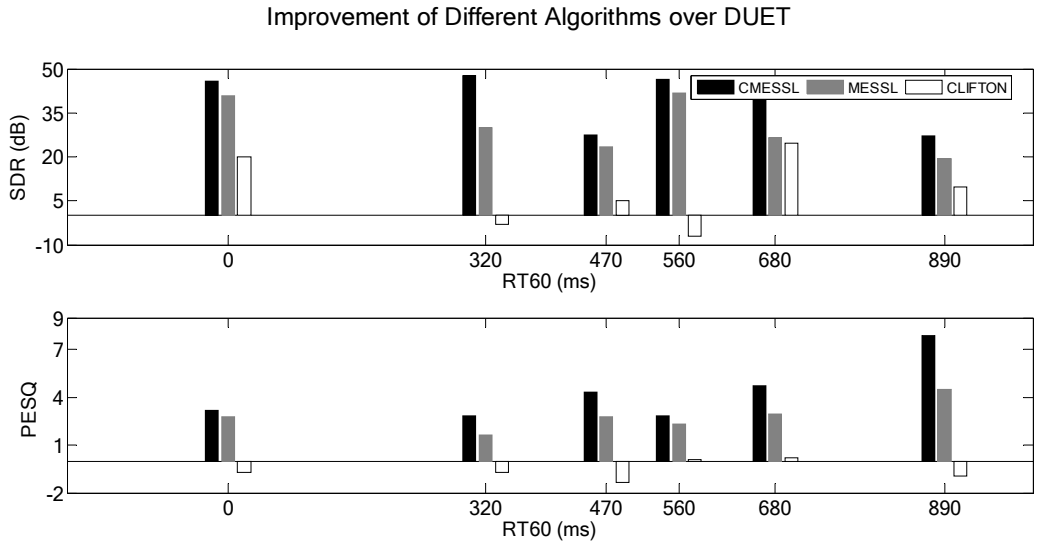

Figure 11: Relative improvement of different models over DUET

B. CASE B: Effect of using Clifton effect with MESSL at small separation angles

In this experiment, we will compare the performance of CMESSL with MESSL at small separation angles. As we do not have RIRs of room S at small separation angles of 5^0 and 10^0, so we will not consider room S in this experiment. We will compare the two algorithms at target-interferer separation angles of 5^0 and 10^0. The results under different acoustic conditions are given in figure 12 to figure 16 below.

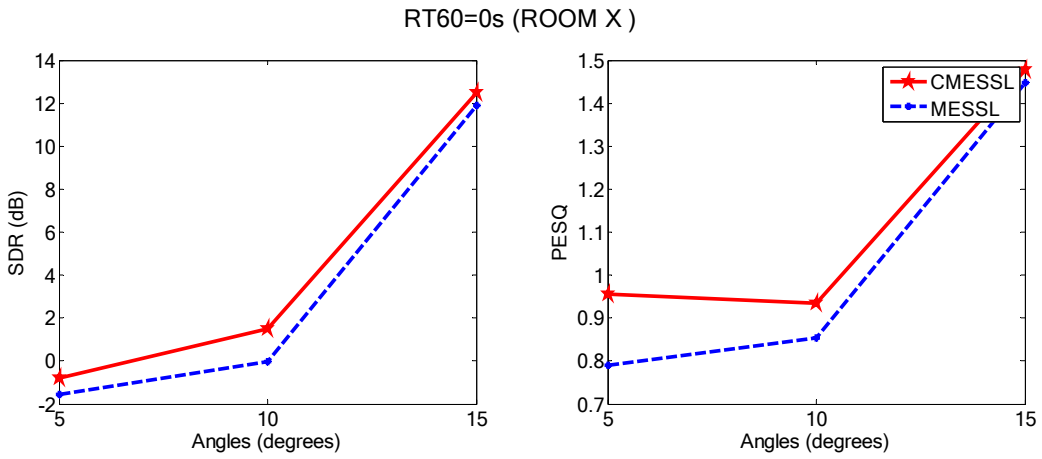

Figure 12: Comparison of MESSL and CMESSL at small separations in Room X (RT60 = 0s)

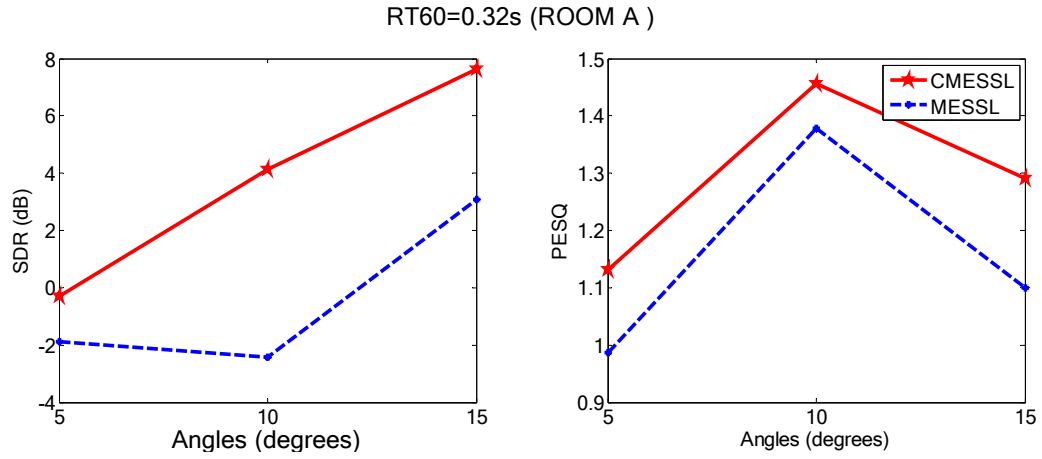

Figure 13: Comparison of MESSL and CMESSL at small separations in Room A (RT60 = 0.32s)

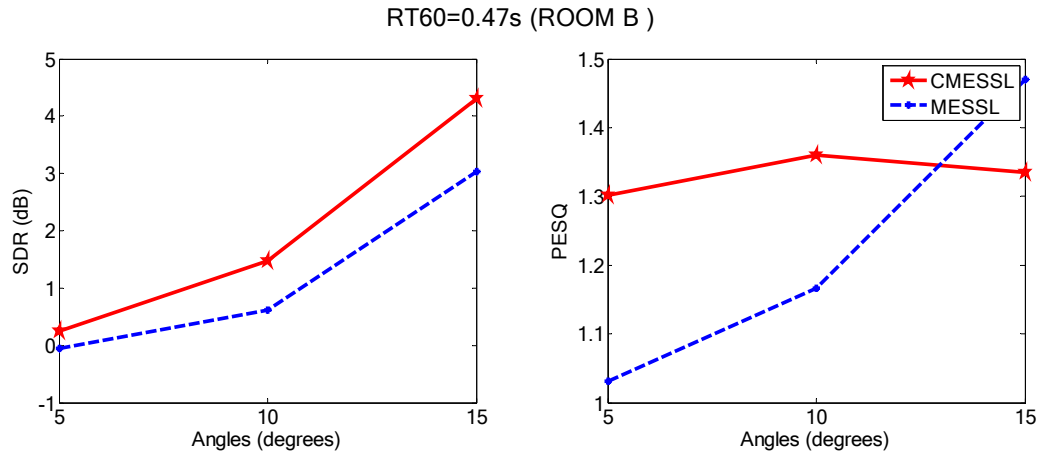

Figure 14: Comparison of MESSL and CMESSL at small separations in Room B (RT60 = 0.47s)

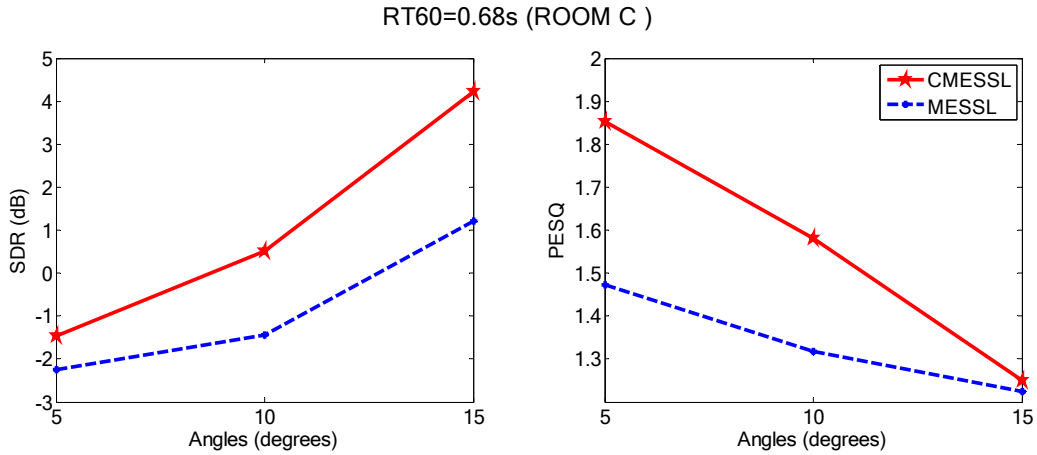

Figure 15: Comparison of MESSL and CMESSL at small separations in Room C (RT60 = 0.68s)

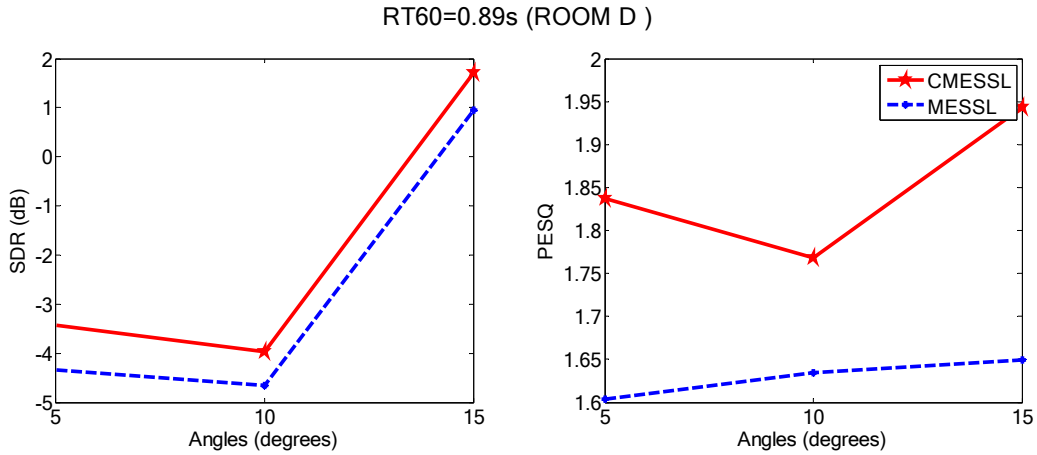

Figure 16: Comparison of MESSL and CMESSL at small separations in Room D (RT60 = 0.89s)

These results show an average improvement of 43% in SDR and 20% in PESQ at 5^0 separations and an improvement of 146 % in SDR and 12% in PESQ at 10^0 separations in all the rooms listed in Table 1 above, excluding the room S.

So, even at smaller separations, our model out-performs MESSL in all types of acoustic conditions proving our supposition that implementing the Clifton effect in dynamic acoustic conditions for a spatial cue based source separation system will improve its separation performance for sources in close proximity.

C. CASE C: Automation Process

As pointed out by [21], the system parameters must adapt themselves with the changing acoustic condition to enable user mobility. So, to enable self-optimization system, we design the automation process, which will work for the rooms in [21]. To adjust the number of channels in gammatone filter bank according to the current acoustic conditions, we need to first find out the RT_{60} of the room. Although

the algorithm in [36] can estimate the room acoustic conditions directly from the individual sources without any prior training, but there are three problems with this algorithm. Firstly, it requires a long duration signal (at least one minute) for accurate RT_{60} estimation, secondly it works fine for individual source but fails for mixtures which contain two or more sources and thirdly it cannot estimate RT_{60}s below 0.2s.

In our proposed model, the values of SDR (Signal to Distortion Ratio) and PESQ (Perceptual evaluation of Speech Quality) of the separated sources are heavily dependent on the room in which the separation is being carried out. So, the SDR and PESQ values can be used as the discriminating features which will decide the class of a test sample.

The classification process consists of two phases; the training phase and the testing phase. In the training phase, the number of channels (N) in the Gammatone Filter Bank is initially set to the value which is common for most of the rooms. So looking at Table 1, we set the initial value of N to six. All training of first classifier is done at this initial value of N.

We will use Bayesian classifier with five classes as the number of rooms in which the automation is sought is five. Each class has an equal prior probability. The SDR and PESQ are recorded at the initial value of N for hundred mixtures, at each angle in the range $\{15^0{:}15^0{:}90^0\}$ in all the rooms. This makes a total of 600 samples of each class. Out of these six hundred samples, 480 samples of each class are used for training and 120 samples for testing.

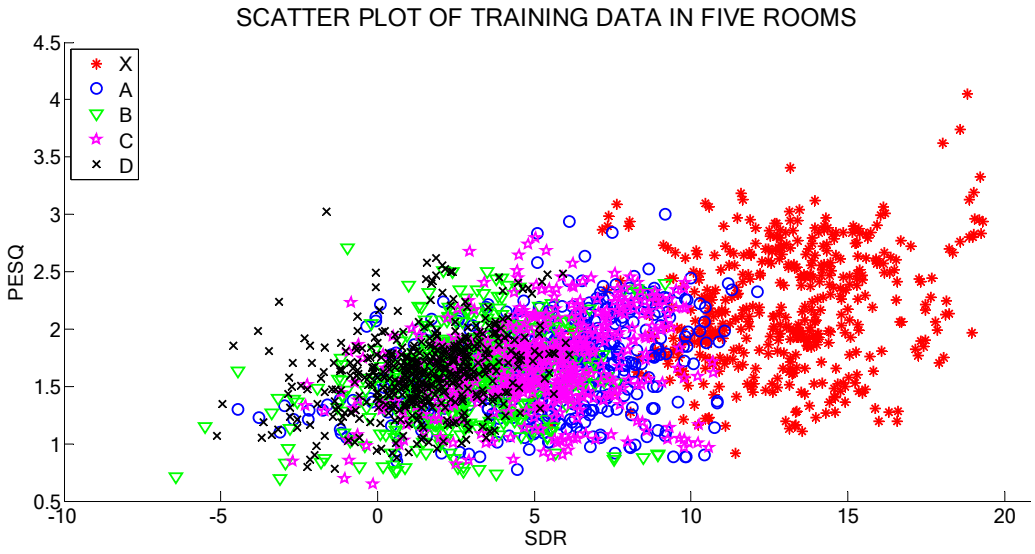

Figure 17: Test samples in five rooms ('X', 'A', 'B', 'C'and 'D')

As can be seen in the scatter plot (Figure 17 above) of the training data in the five rooms in [21], Class X is easily discriminated from all other classes. Any misclassification of test data belonging to class B, C or D to one another does not make any difference as all of them require the same number of channels in the gammatone filter bank as shown in Table 1. However, the samples of class 'A' are uniformly scattered in all other classes and are mainly responsible for lowering the overall accuracy of the classifier. The

accuracy is around 80%. The accuracy can be improved by using advance classification algorithms but it is out of scope of the present paper.

According to the class decided for the test sample, the number of channels (N) in the gammatone filter bank is adjusted. The system carries out this classification process at the output of our proposed model after every 2.5 seconds (this duration depends on the requirement that how fast the system should adjust to the new conditions) by choosing the classifier trained at the current value of N and adapt itself according to the current acoustic conditions. A feedback signal is sent to the gammatone filter bank to adjust its number of channels (N) according to the detected conditions mimicking the process of [25]. It is emphasized again, that this automation process is only applicable for five rooms with two active sources at separation angles in the range $\{15^0{:}15^0{:}90^0\}$.

VII. Discussion and Comparison

Inside a specific acoustic environment, our proposed model uses the Precedence-effect to block the reverberations responsible for the poor performance of generally all spatial cue based source separation systems. The root cause of their poor performance was deformation of spatial cues in reverberant environment, especially when sources were in close proximity. So, these reverberations are blocked to enter our proposed model. Also our proposed model has the ability to adapt itself automatically in dynamic acoustic conditions by utilizing the Clifton effect, where the model parameters are adjusted automatically according to the changing conditions. In its source separation block, our proposed model uses the concept of garbage source (the concept inspired from MESSL) to group up those spatial cues which are coming from the virtual sources created due to reverberations [37], preventing them from messing up with the cues of real sources.

Our proposed model shows improvement over other algorithms listed in this paper for comparison with our proposed algorithm in all kinds of acoustic conditions. It combines best of both, namely, better performance of CLIFTON at smaller separation angles and better performance of MESSL at larger separation angles. Also our model requires a single parameter (the number of channels N in the gammatone filter bank) while CLIFTON model requires two parameters (the inhibitory parameters: the inhibitory time constant (α_p) and the inhibitory gain (G)) to be adjusted with changing acoustic conditions. This parameter in our model is automatically adjusted according to the changing acoustic conditions while it was not possible in CLIFTON. Also our proposed automation process does not require long duration signals, nor it is sensitive to any value of RT_{60} or has limitations about the number of sources in the room. Comparison of different aspects of DUET, CLIFTON and MESSL with our proposed model 'CMESSL' in dynamic acoustic conditions is summarized in Table 3 given below.

Table 3 : Comparison of our proposed model with DUET, CLIFTON and MESSL in dynamic acoustic conditions

Requirements in dynamic acoustic conditions	DUET	CLIFTON	MESSL	CMESSL
Mechanism to deal with reverberations	Not available	Pseudo-Clifton effect	Garbage source	Clifton effect + garbage source
Spatial cue clustering	By maximum likelihood algorithm	On the basis of common azimuth	By EM algorithm	By EM algorithm

Model parameters up gradation	Not supported	Manual	Not supported	Automatic
User mobility	Not supported	Not supported	Not supported	Supported

Table 3 shows that our proposed algorithm uses the Clifton effect to deal with dynamic reverberant conditions, the same psychoacoustic effect used by human brain to deal such situations. The spatial cues are clustered by EM algorithm which provides better source separation as can be verified by the results (higher average SDR and PESQ values) in all kinds of acoustic conditions ranging from anechoic to highly echoic and the system supports user mobility, the most highlighted feature of our proposed model not supported by DUET, CLIFTON or MESSL.

Better separation results over other competing algorithms (namely DUET, MESSL and CLIFTON) in all types of acoustic conditions, blind extraction of acoustic conditions from the output quality of separated sources, adjustment of model parameters automatically according to the existing conditions and improved performance for closely located sources are the key strengths of our proposed model which make it stand out among the competitors.

VIII. Conclusion and Future Work

Our model gives general idea of improvement under dynamic reverberant conditions in spatial source separation systems using the Clifton effect considering only the two source case. However, its results are generally applicable for more number of sources but due to the reduction in sparseness and increased ambiguity in resolving the similar spatial cues for closely located sources, there will be an observable decline in separation performance. The automation process makes this model useful for source separation when the user is roaming in dynamic acoustic conditions. Also, our proposed algorithm shows improvement over other spatial source separation algorithms for sources in close proximity due to effectively suppressing the echoes which were responsible for distorting the spatial cues used by these systems for segregation of sources. The task of source separation utilizing spatial cues, only two mixtures, and automatic adaptation to acoustic condition make our proposed model an excellent addition to CASA based approaches of source separation. In 2008, voice processing chip based on CASA technology was announced by Audience Inc. [38] to improve mobile phone call quality. This chip achieved noise suppression of 25 dB. This chip if equipped with our proposed algorithm will not only suppress the noise but also the reverberations entering the phone when user is standing or roaming in enclosed area where reverberations and noise are among the key factors effecting the voice quality.

Future work will focus on the improvement of accuracy of automation process in scenarios where the user is allowed to roam freely both inside the room and from one room to another. Another interesting application would be to use our proposed algorithm in assisted protection headphones for separating noise from the required signal. These headphones are designed to prevent hearing loss to construction workers, workers of factories, musicians and workers of night clubs who are exposed to occupational noise for long durations [39]. The headphone model proposed in [40] suppresses both the signal of interest and noise without any discrimination between them when their level increases beyond the pre-set threshold. However, using our proposed model in these headphones, the user can smartly turn off the noise without suppressing the signal of interest.

Funding

This project is funded by Higher Education Commission (HEC), Pakistan, under project no. 6330/KPK/NRPU/R&D/HEC/2016.

Acknowledgements

We would like to thank Christopher Hummersone and Michael I. Mandel for sharing their codes and our fellow researchers Fazli-Hadi and Faiq Ahmad Khan for their help in simulations.

References

[1]. Ganesh R. NAIK.: "Measure of quality of source separation for suband super-gaussian audio mixtures", INFORMATICA, Vol. 23, No. 4, pp 581–599 , Vilnius University, 2012.

[2]. Jacob Benesty, Shoji Makino, Jingdong Chen.: "Speech enhancement", ISBN 3-540-24039-X, Springer, 2005.

[3]. Edward J. Devinney Jr., Manish Sharma, Chris Keyser, Rainer Rothacker.: "User validation For Information System Access And Transaction Processing", United states, Patent Pub. No.: US 2003/0046083 A1, Mar. 6, 2003.

[4]. Cherry, E.C.: "On human communication" (MIT Press, Cambridge, MA, 1957).

[5]. D.L.Wang and G.J. Brown.: "Computational auditory scene analysis: principles, algorithms, and applications" (Wiley, Hobboken, NJ, USA 2006).

[6]. Anna K. N´ab˘elek and Pauline K. Robinson.: "Monaural and binaural speech perception in reverberation for listeners of various ages", Journal of the Acoustical Society of America, 1982, 71(5), pp1242–1248.

[7]. Bulent Bilgehan and Amr Abdelbari.: "Fast detection and DOA estimation of the unknown wideband signal sources", International Journal of Communication Systems, April 2019, Vol 32, issue 11.

[8]. Shiann-ShiunJeng, Hsin-Piao Lin and Chan-Wan Tsung.: "Experimental studies of direction of arrivals using a smart antenna testbed in wireless communication systems", International Journal of Communication Systems, March 2003, Vol 16, issue 3.

[9]. Michael I. Mandel, Ron J. Weiss, Daniel P. W. Ellis.: "Model-based expectation-maximization source separation and localization", IEEE transactions on audio, speech and language processing, February 2010, Vol. 18, No. 2.

[10]. Z.Y. Zohny, S.M. Naqvi and J.A. Chambers.: "Modeling inter-aural level and phase cues with student´s T-distribution for robust clustering in MESSL", Proc. 19th International conference on DSP, Aug 2014.

[11]. Z.Y. Zohny.: "Robust variational Bayesian clustering for underdetermined speech separation". Ph.D. thesis, Loughborough University, UK, 2016.

[12]. Muhammad Salman Khan, Syed Mohsen Naqvi, Ata-ur-Rehman, and Jonathon Chambers.: "Video-aided model-based source separation in real reverberant rooms", IEEE transactions on audio, speech and language processing, September 2013, Vol. 21, No. 9, pp 1900-1911.

[13]. M. S. Khan, A.-Rehman, S. M. Naqvi, and J. A. Chambers.: "Convolutive speech separation by combining probabilistic models employing the interaural spatial cues and properties of the room assisted by vision", Proc. 9th IMA Mathematics in Signal Processing, Birmingham, UK, 2012.

[14]. M. S. Khan, S. M. Naqvi, and J. A. Chambers.: "Two-stage audio-visual speech dereverberation and separation based on models of the interaural spatial cues and spatial covariance", Proc. IEEE DSP, Santorini, Greece, 2013.

[15]. S. M. Naqvi, W. Wang, M. S. Khan, M. Barnard, and J. A. Chambers.: "Multimodal (audio-visual) source separation exploiting multi-speaker tracking, robust beamforming, and time-frequency masking", IET Signal Processing, 2012, vol. 6, no. 5, pp. 466-477.

[16]. S. M. Naqvi, M. S. Khan, Q. Liu, W. Wang, and J. A. Chambers.: "Multimodal blind source separation with a circular microphone array and robust beamforming", Proc. EUSIPCO, Barcelona, Spain, 2011.

[17]. Xueliang Zhang, DeLiang Wang.: "Deep learning based binaural speech separation in reverberant environments", IEEE/ACM transactions on audio, speech and language processing, May 2017, Vol. 25, No. 5.

[18]. ZhongQiuWang and DeLiangWang.: "Combining spectral and spatial features for deep learning based blind speaker separation", IEEE/ACM transaction on audio, speech and language processing, Feb 2019, Vol 27, Issue 2.

[19]. Wallach, Newman, and Rosenzweig.: "The precedence effect in sound localization", The American Journal of Psychology, 1949, 62, 3, pp 315–336.

[20]. Ardiansyah Musa and Gde Dharma Hugraha.: 'A decision tree based NLOS detection method for the UWB indoor location tracking accuracy improvement", International Journal of Communication Systems, June 2019.

[21]. Christopher Hummersone.: "A psychoacoustic engineering approach to machine sound source separation in reverberant environments", Ph.D. thesis, University of Surrey, February 2011.

[22]. M. S. Khan, S. M. Naqvi, and J. A. Chambers.: "Speech separation with dereverberation based pre-processing incorporating visual cues", Proc. 2nd International Workshop on Machine Hearing In Multisource Environments (CHIME), Vancouver, Canada, 2013.

[23]. M. S. Khan, S. M. Naqvi, and J. A. Chambers.: "A new cascaded spectral subtraction approach for binaural speech dereverberation and its application in source separation", Proc. IEEE ICASSP, Vancouver, Canada, 2013.

[24]. Kalle J. Palomaki, Guy J. brown and DeLiang Wang.: “A binaural processor for missing data speech recognition in the presence of noise and small room reverberation”, Elsevier speech communication, March 2004, 43, pp 361-378.

[25]. Litovsky, Rakerd, Yin and Hartmann.: “Psychophysical and physiological evidence for a precedence effect in the median sagittal plane”, Journal of Neurophysiology, 1997, 77, 4, pp 2223–2226.

[26]. Cooke.:“Modeling auditory processing and organization”, Ph.D. thesis, University of Sheffield, 1991.

[27]. Moore.: “An introduction to the psychology of hearing” (London, Academic Press, fifth edition. 2004).

[28]. P. Aarabi and A. Mahdavi.: “The relation between speech segment selectivity and time-delay estimation accuracy”, Proc. IEEE conference on acoustics, speech, signal processing, May 2002.

[29]. B. Shinn-Cunningham, N. Kopco and T. Martin.: “Localizing nearby sound sources in a classroom: binaural room impulse responses”, Journal of Acoustical Society of America, 2005, Vol. 117, No. 5, pp 3100–3115.

[30]. Ozgur Yilmaz and Scott Rickard.: “Blind separation of speech mixtures via time frequency masking”, IEEE transactions on signal processing, July 2004, Vol. 52, No. 7.

[31]. Shoko Araki, Shoji Makino, Hiroshi Swadaand Ryo Mukai.: “Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA”, EUSIPCO, 2004, pp 899-905.

[32]. “DAPRA TIMIT acoustic phonetic continuous speech corpus”, http://www.ldc.upenn.edu/Catalog/LDC93S1.html, accessed 10th January 2019.

[33]. Emmanuel Vincent, RemiGribonval and Cedric Fevotte.: “Performance measurement in blind audio source separation”, IEEE transactions on audio, speech and language processing, July 2006, Vol. 14, No. 4, pp 1462-1469.

[34]. ‘PESQ files’, https://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en, accessed 10th January 2019.

[35]. Scott Rickard.: “The DUET blind source separation algorithm”, University College, Dublin (Springer 2007), pp 217-241.

[36]. Heinrich W. Löllmann, EmreYilmaz, Marco Jeub and Peter Vary.: "An improved algorithm for blind reverberation time estimation", Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv, Israel, August 2010.

[37]. dr.ir. Emanu□el A.P. Habets.: “Room impulse response generator”. Report by ‘dereverberation.org’, September 20, 2010.

[38]. Press release: Audience introduces industry-first voice processor based on human hearing system and begins sampling to mobile handset manufacturers, http://embedded-computing.com/news/audience-sampling-mobile-handset-manufacturers/

[39]. L García, Let al.: "Valencia's cathedral church bell acoustics impact on the hearing abilities of bell ringers", International journal of environmental research and public health 16 (9), 1564, 2019.

[40]. L Parra, M Torres, J Lloret, A Campos, I Bosh.: "Assisted protection headphone proposal to prevent chronic exposure to percussion instruments on musicians", Journal of healthcare engineering, 2018.

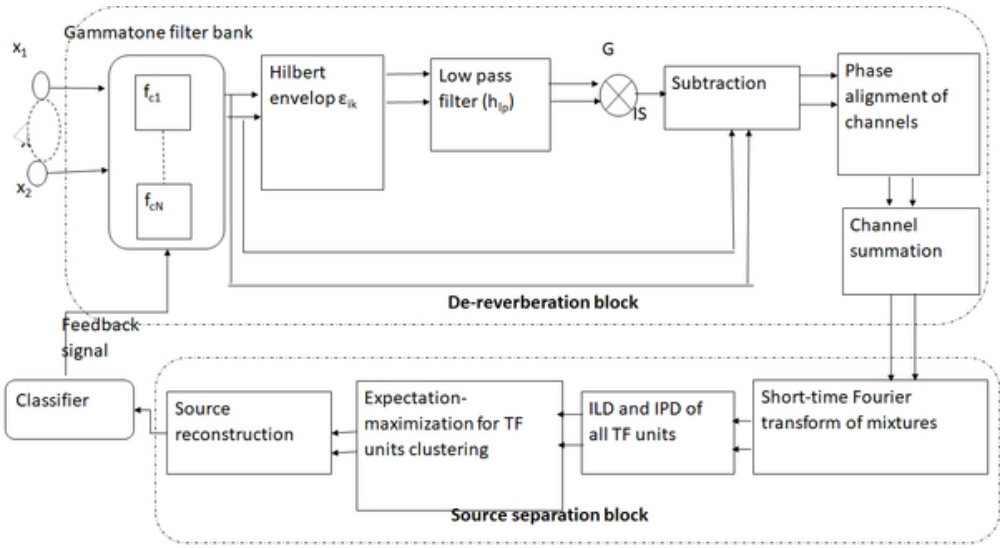

Figure 1: Block diagram of CMESSL, showing the application of the Clifton-effect on speech mixtures followed by source separation process

50x27mm (300 x 300 DPI)

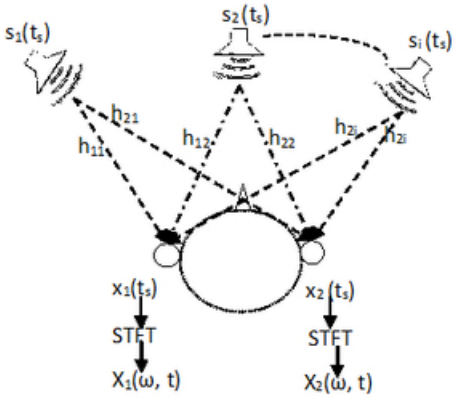

Figure 2: Signal notations. The two mixture signals are transformed to the time frequency domain X1(ω, t) and X2(ω, t) by STFT.

26x12mm (600 x 600 DPI)

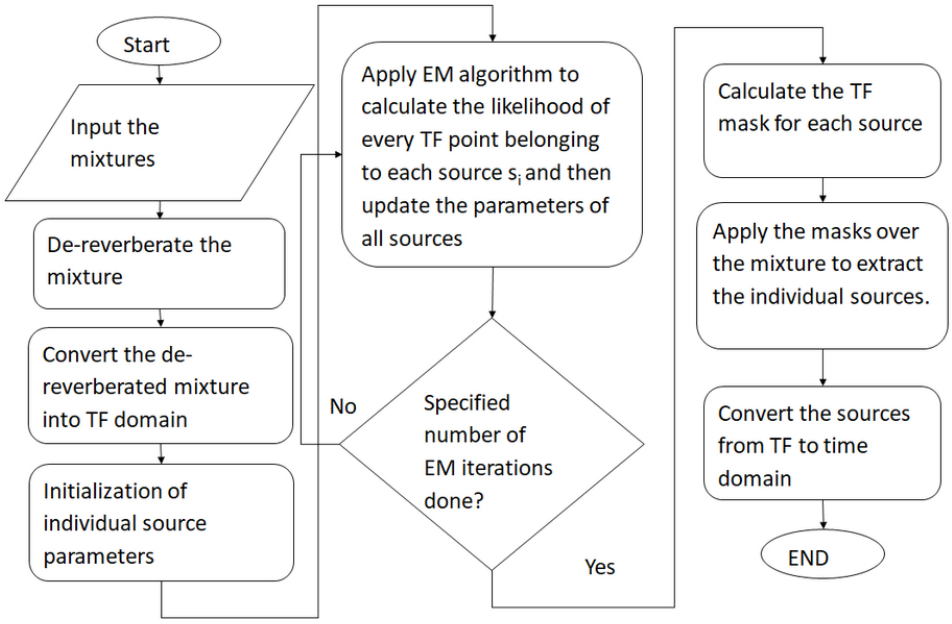

Figure 3: Flow Chart of our proposed algorithm

40x24mm (600 x 600 DPI)

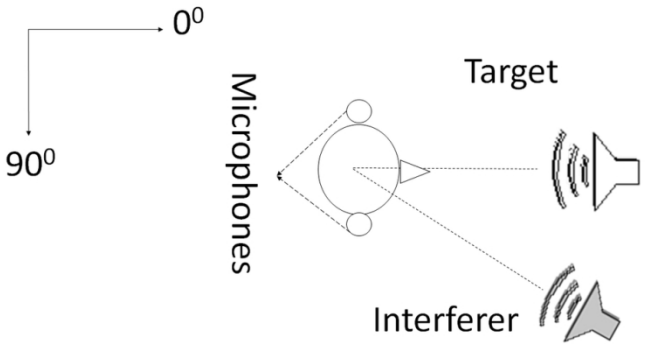

Figure 4: The room layout showing one of the approximate positions of the sources and the sensors.

61x38mm (600 x 600 DPI)

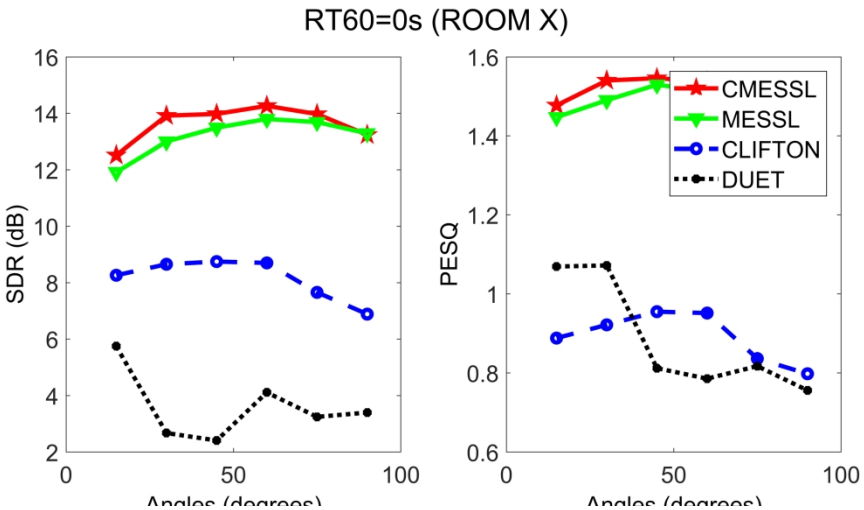

Figure 5: Comparison of algorithms in Room X with RT60 of 0ms.

242x124mm (300 x 300 DPI)

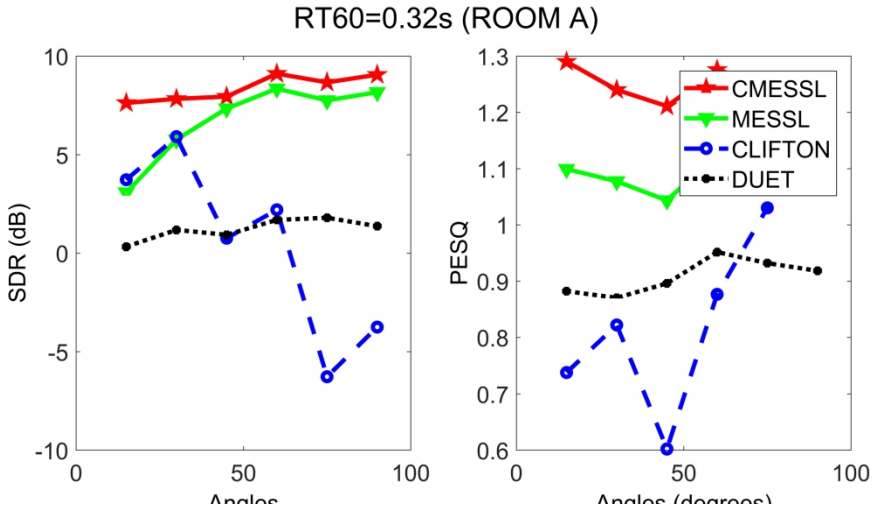

Figure 6: Comparison of algorithms in Room A with RT60 of 320ms.

242x124mm (300 x 300 DPI)

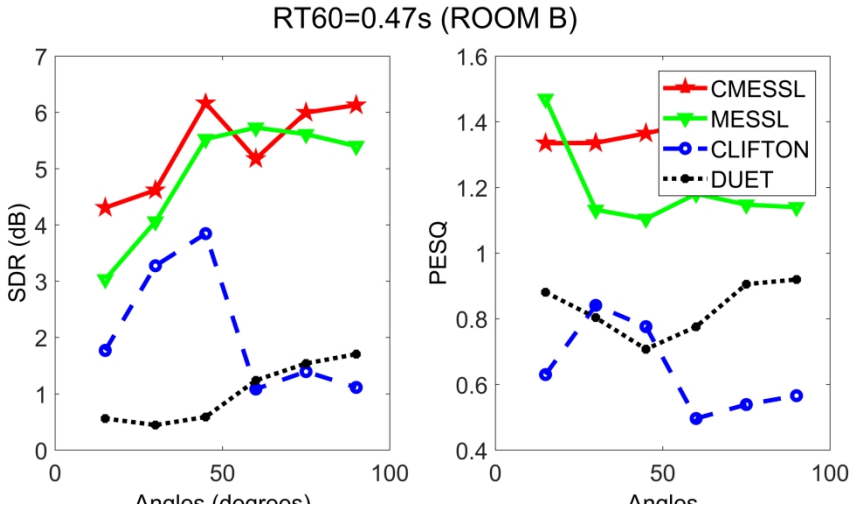

Figure 7: Comparison of algorithms in Room B with RT60 of 470ms.

242x124mm (300 x 300 DPI)

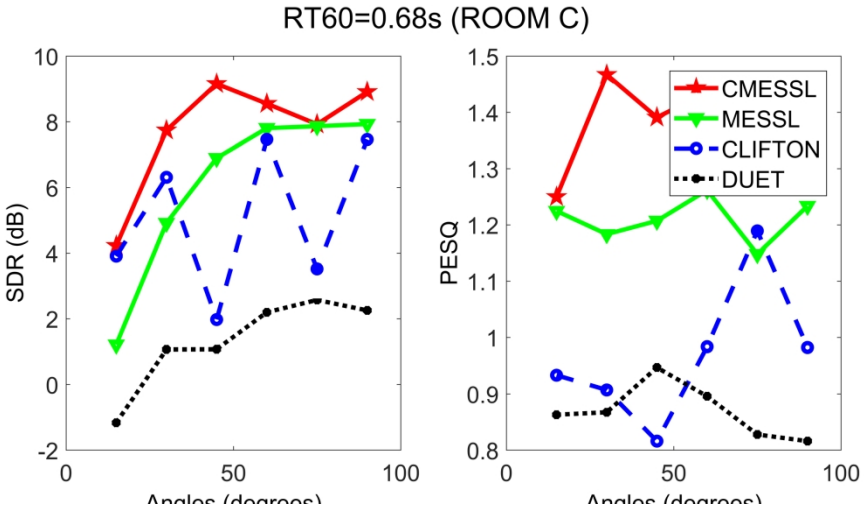

Figure 8: Comparison of algorithms in Room C with RT60 of 680ms.

242x124mm (300 x 300 DPI)

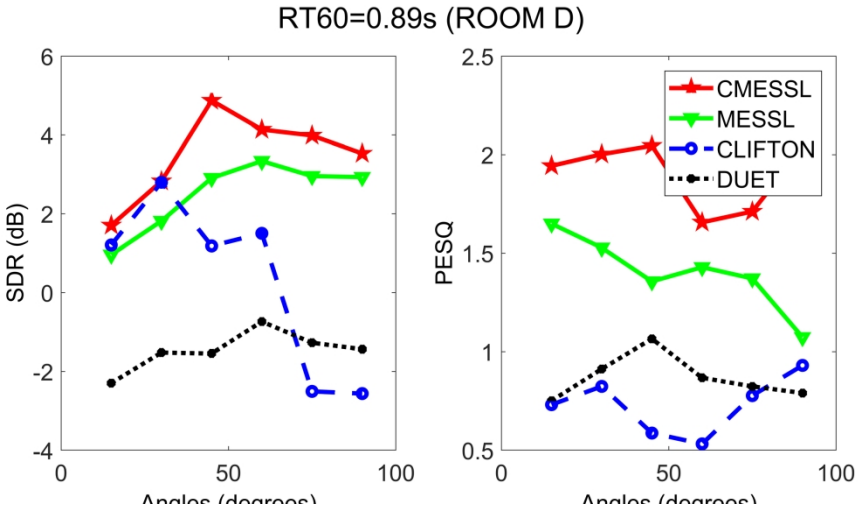

Figure 9: Comparison of algorithms in Room D with RT60 of 890ms.

242x124mm (300 x 300 DPI)

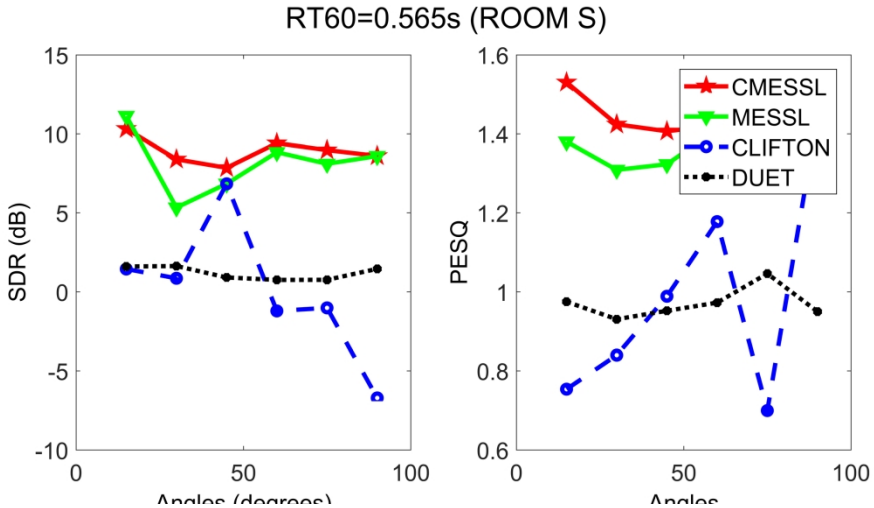

Figure 10: Comparison of algorithms in Room S with RT60 of 565ms.

242x124mm (300 x 300 DPI)

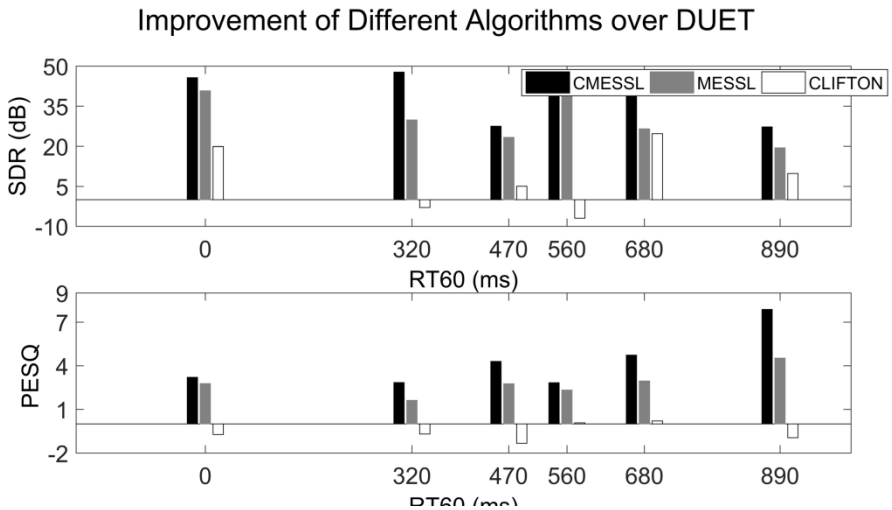

Figure 11: Relative improvement of different models over DUET

242x124mm (300 x 300 DPI)

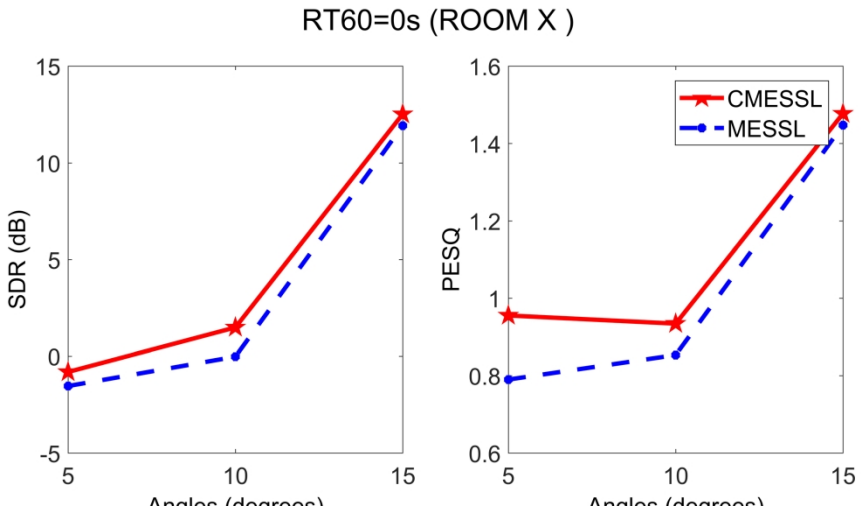

Figure 12: Comparison of MESSL and CMESSL at small separations in Room X (RT60 = 0s)

242x124mm (300 x 300 DPI)

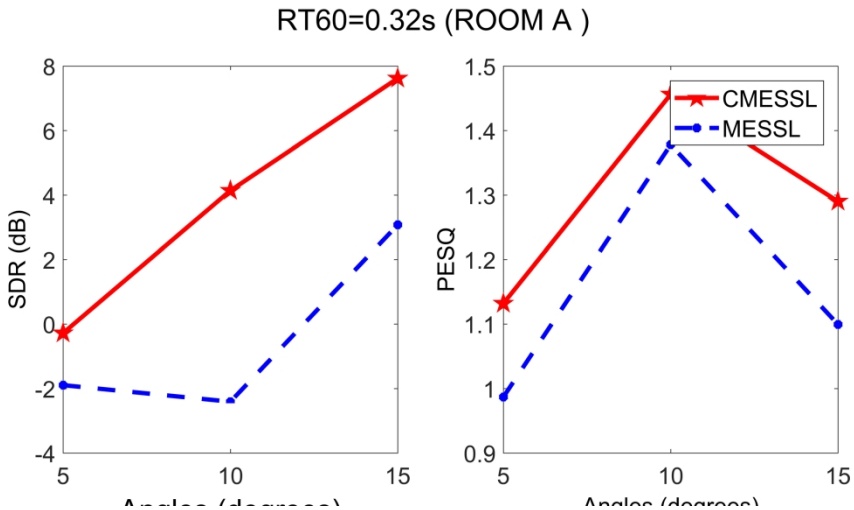

Figure 13: Comparison of MESSL and CMESSL at small separations in Room A (RT60 = 0.32s)

242x124mm (300 x 300 DPI)

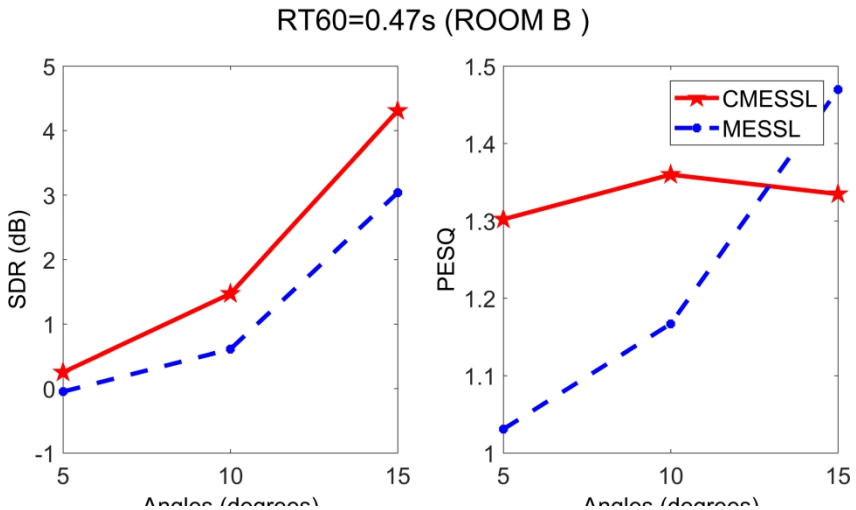

Figure 14: Comparison of MESSL and CMESSL at small separations in Room B (RT60 = 0.47s)

242x124mm (300 x 300 DPI)

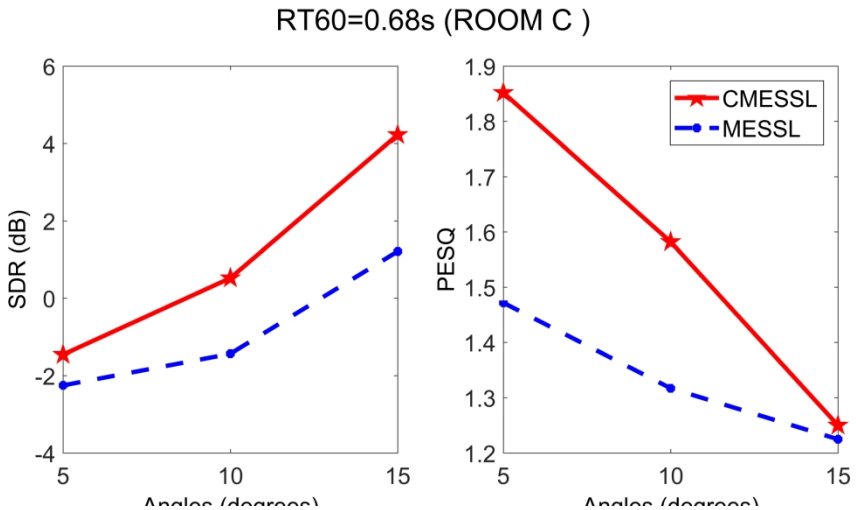

Figure 15: Comparison of MESSL and CMESSL at small separations in Room C (RT60 = 0.68s)

242x124mm (300 x 300 DPI)

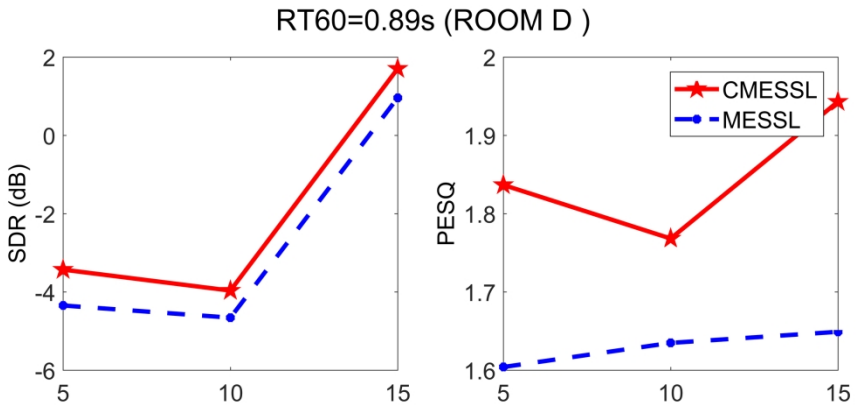

Figure 16: Comparison of MESSL and CMESSL at small separations in Room D (RT60 = 0.89s)

242x100mm (300 x 300 DPI)

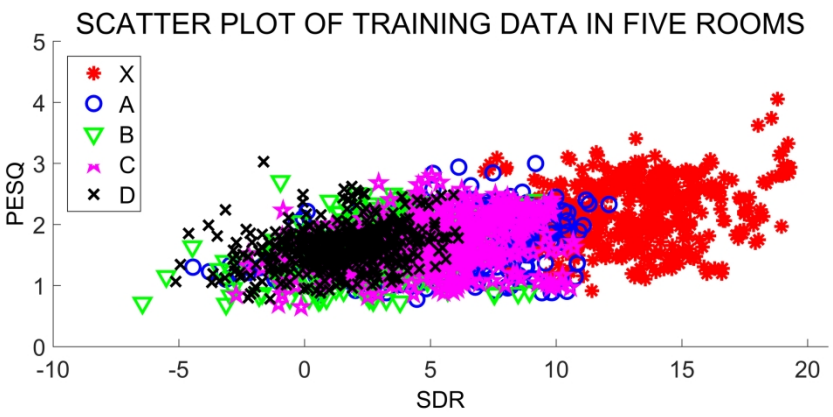

Figure 17: Test samples in five rooms ('X', 'A', 'B', 'C'and 'D')

242x100mm (300 x 300 DPI)