Multi-microphone recording speech enhancement approach based on pre-processing followed by multi-channel method

HÉLA KHAZRI, MOHAMED ANOUAR BEN MESSAOUD, AICHA BOUZID
National School of Engineers of Tunis, University of Tunis El Manar
TUNISIA
khazri.hella@gmail.com, anouar.benmessaoud@yahoo.fr, bouzidacha@yahoo.fr

Abstract: - In this paper, we propose an efficient multi-channel speech enhancement approach, based on the idea of adding a pre-treatment preceding the speech enhancement via a multi-channel method. This approach consists at first step in applying mono-channel speech enhancement method to process each noisy speech signal independently and then applying a multi-channel method based on the delay estimation and the blind Speech Separation in order to obtain the enhanced speech. Our idea is to apply a different class of mono-channel method in order to compare between them and to find the best combination that can remove a maximum noise without introducing artifacts. We resort the use of two classes of algorithms: the spectral subtraction and the statistical model based methods. In order to evaluate our proposed approach, we have compared it with our multi-channel speech enhancement method without a preprocessing. Our evaluation that was performed on a number of records corrupted by different types of noise like white, Car and babble shows that our proposed approach provides a higher noise reduction and a lower signal distortion.

Key-Words: - Speech enhancement, Mono-channel Speech Separation, Multi-channel Speech Separation, Delay Estimation, Spectral Subtraction, Statistical Model Based Methods

1 Introduction

Recently, the real-world environment is always degraded by additive background noise such as hands-free speech recognition system, and teleconferencing systems**.** Speech enhancement systems are often used in such situation to improve the perceptual quality, intelligibility of speech degree of listener fatigue by minimizing the effect of noise for an increasing number of speech applications.

Generally, speech enhancement systems can be divided into two general groups: the first group is based on mono-channel techniques such as wavelet transforms, spectral subtraction algorithms [1], wiener filtering [2], statistical-model-based-Methods [3] and Subspace algorithms [4] and the second group is based on multi-channel techniques. First of all, comparing with the human auditory system, the multi-channel speech enhancement systems represent also the more realistic system, due to its spatial filtering capability of suppressing the interfering signals arriving from directions other than the specified look-direction.

Multi-microphone speech enhancement algorithms take advantage of the availability of multiple signal input to our system. These noise reduction algorithms prove in addition its performance in reducing speech distortion and musical noise since they may utilize both the temporal and the spatial domain. That‘s why in the last few decades, multi-microphone speech enhancement algorithms have attracted a great deal of interest. There are various studies on multi-channel speech enhancement; in particular, the array and adaptive beamformer (ABF) [5]–[6], the delay-and-sum (DS) [7] and the Blind source separation (BSS) [8]–[9].

The ABF is the most conventionally used microphone arrays for source segregation and noise reduction. However, the adaptive beamformer requires a speech break interval, and *a* priori information. These requirements are due to the fact that the ABF is based on supervised adaptive filtering, which significantly limits the applicability of ABF to source separation in practical applications. Indeed, the adaptive beamformer cannot work well when the interfering signal is non-stationary noise.

The BSS is a method to determining original source signals using only mixed speech observed in each input channel. In particular, BSS based on independent component analysis (ICA) is applied [10]. Indeed, the conventional ICA could work particularly in multi-speaker separation, but such a mixing condition is very rare and unrealistic; real noises are often widespread sources. In this paper, we mainly deal with generalized noise that cannot be regarded as a point source. Moreover, we assume

this noise to be non-stationary noise that arises in many acoustical environments; however, ABF could not treat this noise well. Although ICA is not influenced by non-stationarity of signals unlike ABF, this is still a very challenging task that conventional ICA-based BSS could hardly address because ICA cannot separate widespread sources.
In order to improve the performance of our approach, a technique combines the improved delay-and-sum approach and a modified version of spectral subtraction mono-channel has been proposed.
Generally, the mono-channel and the multi-channel algorithms both of them aim to improve the overall speech signals quality and enhancing intelligibility. Or the main mono-channel speech enhancement's disadvantage is that they fail to completely eliminate noise and the musical noise's generation. So, our real goal is to find a compromise between minimization of the distortion introduced into the signal and maximization the reduction of the noise. That is why the choice of the appropriate denoising method is very important and it depends mainly on the model of the signals recorded and noise's type.

In this work, we propose a multi-channel speech enhancement approach, based on adding a pre-processing preceding the speech enhancement via a multi-channel method. Our basic idea is to combine a mono-channel speech enhancement method that treats each channel independently. Then, these enhanced speech obtained are processed by a multi-channel speech enhancement method based on the delay estimation.
Our goal is to apply a different mono-channel method in order to compare between them and to find the best combination that can remove a maximum noise without speech distortion.
We used two different mono-channel methods as a pre-processing phase: the geometric approach of the spectral subtraction and the estimators of the magnitude-Squared Spectrum. The proposed method is tested on noisy speech under various noise conditions including white, babble and volvo. Objective and subjective results shows that the system based on this approach has significant improvement over this recent method.
The rest of the paper is organized as follows. In section 2, we describe our proposed multi-channel speech enhancement approach by giving a detailed overview of the methods used in it. The experimental results of our method under a variety of real noisy environments are given in section 3. Finally, we draw conclusions in section 4.

2 Our proposed approach

Aimed to improve speech perceptual quality and intelligibility in the acoustical environment, a speech enhancement system is proposed in this section, which consists of two general stages: First, a pretreatment is applied by processing a mono-channel speech enhancement method to each noisy speech signal independently. Secondly, a multi-channel method based on the delay estimation and the blind Speech Separation is applied. Recently, in many speech communication applications multi-microphone speech enhancement techniques can be used instead of single-microphone speech enhancement techniques. A well-known there are many single-microphone techniques that prove their performance in the speech enhancement's field. That's why we can take privilege of their performance by combining a mono-microphone and multi-microphone speech enhancement methods in order to simultaneously minimize speech distortion and maximize noise reduction.
In the following diagram, we present our multi-microphone noise reduction approach's generalized scheme that consists an adding a pretreatment preceding speech enhancement via a multichannel method:

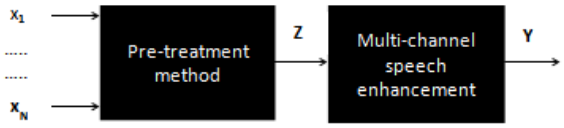

Figure 1.Our proposed approach

In this paper, a novel blind multi-channel speech enhancement system is proposed that combines pre-processing method and blind multi-microphone noise reduction method. We aimed to develop a flexible approach that combines the strengths of multi-channel enhancement techniques with noise reduction algorithms. Fig.1 shows the overall framework of the proposed system. It consists of two parts: Pre-processing method and multi-microphone speech enhancement method.

In the first stage, noise reduction, is performed. After applying single microphone noise reduction algorithm to the mixture in each channel individually, the geometric approach to spectral subtraction or the statistical estimators of the magnitude-squared spectrum are used to remove some types of noise. In the second stage, multi-channel speech enhancement techniques based on time-delay estimation, such as delay-and-sum, delay-and-feature-domain-sum and phase-error based filtering are applying on the processed signals

to significantly improve the perceptual speech's quality and intelligibility.

2.1 Mono-channel speech enhancement methods

We aim to propose an approach that provides a higher noise reduction and a lower signal distortion in the context of mono-channel speech enhancement. Speech signals can be first enhanced by pre-processing so that recognition performs better on successive steps. It is understood that pre-processing is performed at the signal level. The following list summarizes some of these methods used as the first enhancement step for speech signals in our proposed approach:

- The geometric approach of the spectral subtraction
- The estimators of the magnitude-Squared Spectrum

Spectral subtraction algorithms are a reference algorithm for noise reduction. In 1979, Boll proposed an algorithm that operates in the frequency domain using spectral changes. Boll's algorithm becomes one of the earliest and the most popular speech enhancement method. It is simple and easy to implement it but it suffers from musical noise and signal distortion. So, different derived methods from it are proposed such as spectral subtraction with over subtraction factor, nonlinear spectral subtraction, multiband spectral subtraction, minimum mean square error spectral subtraction, selective spectral subtraction, spectral subtraction based on perceptual properties and the geometric approach to spectral subtraction (GA) [11].

In 2008, Lu and Loizou proposed that the noisy spectrum at frequency ω can be represented geometrically in the complex plane as the sum of two complex numbers: the clean signal spectrum and the noise spectrum at this frequency [12]. In the next figure we present the geometric approach of the spectral subtraction's implementation.

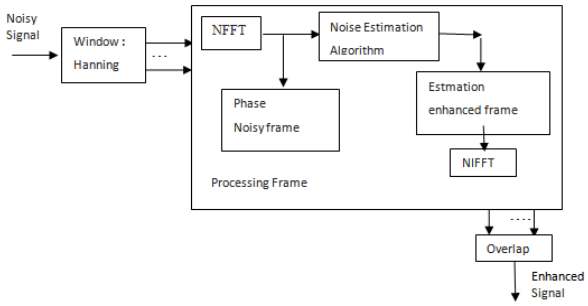

Figure 2.The GA's implementation

Among these modified method the geometric approach of the spectral subtraction proves its performance in noise reduction without affecting the speech signal quality.

In 1984, a basic estimator was proposed by Ephraim and Malah based on an estimation of short-term spectrum in the least squares sense named Minimum Mean Square Error-Short-Term Spectral Amplitude (MMSE-STSA) [13]–[14]. In 2011, Lu and Loizou [15] propose the magnitude-Squared Spectrum estimators which based on the assumption that the magnitude-squared spectrum of the speech signal is the sum of the (clean) signal and the noise magnitude-squared spectra.

The estimators of the magnitude-Squared Spectrum [15] can be classified into two categories:

- Hard masking estimators:

  ✓ Maximum a posteriori estimator (MAP) is given as follows:

$$y_k^2 = \begin{cases} X_k^2, if\, \sigma_s^2(k) \geq \sigma_b^2(k) \\ 0, if\, \sigma_s^2(k) < \sigma_b^2(k) \end{cases} \quad (1)$$

  where: Y_k^2 , X_k^2 $et B_k^2$ are respectively the signal magnitude-squared spectrum of the estimated, noisy speech and the noise.

  and: $\sigma_s^2(k) \equiv E\{S_k^2(k)\}, \sigma_b^2(k) \equiv E\{B_k^2(k)\}$

  ✓ Minimum Mean Square Error estimator (MMSE): Reference [15] shows, there are two derivations Of the MMSE estimator: MMSE-SPZC and MMSE-SPZC-SNRU

$$Y_k^2 = \begin{cases} \left(\frac{1}{\nu_k} - \frac{1}{e^{\nu_k - 1}}\right) X_k^2, if\, \sigma_s^2(k) \neq \sigma_b^2(k) \\ \frac{1}{2} X_k^2, if\, \sigma_s^2(k) = \sigma_b^2(k) \end{cases} \quad (2)$$

  where:

$$\chi_k \equiv \frac{1-\xi_k}{\xi_k}\gamma_k$$

$$\xi_k \equiv \frac{\sigma_s^2(k)}{\sigma_b^2(k)}$$

$$\gamma_k \equiv \frac{X_k^2}{\sigma_b^2(k)}$$

- Soft masking estimators: In [15], the proposed method shows that the estimators based on the soft masking using the uncertainty of SNR can be divided into:

  ✓ Estimator incorporating a Priori SNR uncertainty (SMPR) is given by:

$$Y_k^2 = \frac{\xi_k}{\xi_k+\theta}.X_k^2 \quad (3)$$

  note that ξ_k is the Priori SNR

  ✓ Estimator incorporating a Posterior SNR uncertainty (SMPO) is given as follow:

$$Y_k^2 = \begin{cases} \frac{e^{\frac{v_k}{\theta+1}}-1}{e^{v_k}-1}.X_k^2, if\sigma_s^2(k) \neq \sigma_b^2(k) \\ \left(\frac{1}{\theta+1}\right).X_k^2, if\sigma_s^2(k) = \sigma_b^2(k) \end{cases} \quad (4)$$

These estimators become the most popular mono-channel speech enhancement algorithm which shows better performance than the other algorithms since their satisfactory results in terms of noise reduction and the speech distortion.

2.2 Multi-channel speech enhancement methods

Recently, multi-channel speech enhancement methods can be classified into two categories:

- Conventional multi-microphone speech enhancement methods
- Blind multi-microphone speech enhancement methods

Conventional multi-channel speech enhancement methods require certain a priori knowledge about the signals and the environment.

Blind multi-channel speech enhancement methods don't require any knowledge about the signals and no other information about the signal distortion on the transfer paths from the sources to the sensors is available. The only a priori knowledge about them is the statistical independence's signals.

According to the nature of the problem to solve, we investigate two multi-microphone speech enhancement methods [16] that have gained a lot of attention:

- Delay and Sumbeamformer
- Phase-errorbasedfiltering

Delay-and-sum beamformer and Phase-error based filtering which represent two multi-channel speech enhancement methods depend on two principles techniques:

Blind source separation (BSS): The speech signal gets distorted and mixed when transmitted from sources to be recorded by a set of microphones in a multi-channel environment. Humans have the ability to recognize a specific voice among a din of conversations and background noise, known as the "cocktail party effect". The problem of Blind source separation (BSS) is the separation of a set of source signals from a set of mixed signals, without any information about the source signals and the mixing process. BSS consists of recovering unknown signals or "sources" from their several observed mixtures. There are many solutions of the BSS problem: Independent Component Analysis (ICA), Independent Factor Analysis (IFA) [17].

Time Delay Estimation (TDE): The problem of estimating the delay between signals recorded by a set of microphones in a noisy environment has be provoked in a variety of applications such as microphone array processing systems and speech enhancement. Reference [18] shows that there are many algorithms to estimate the time delay: Cross-correlation (CC) method, Phase transform (PHAT) method, Maximum likelihood (ML) method, Average square difference function (ASDF) method, least mean square (LMS) adaptive filter method. Among these algorithms above, we turn to account these two techniques:

- The cross-correlation method: this technique computes the cross correlation function between the received signals and considers the maximum peak in the output as the estimated time delay.
- The Phase Transform (PHAT): In 1976, the generalized cross-correlation (GCC) is proposed by Knapp and Carter [19]. The Phase Transform is a GCC procedure that uses a weighting functions after the cross correlation to improve the time delay estimation.

Delay-and-sum beamformer (DS) is the simplest and the most popular beamforming algorithm. It aims to appropriately compensate signal delay for each channel before summing them. The noise in each microphone tends to statistically cancel each other. To precede it, the proper time delay rated the time-difference of arrival (TDOA) has to be estimated using a Time Delay Estimation (TDE) technique. So to do it a one channel has to be chosen as a reference.

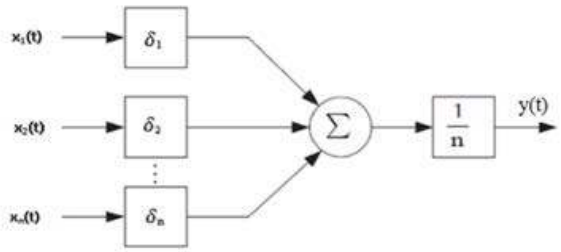

Figure 3.Delay and Sum beamformer block diagram

The enhanced signal y(t) resulting from the delay and sum method is described by the following equation:

$$y(t) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \delta_i) \quad (5)$$

where n is the number of channels and x_i(t) is the signal received by the i[th] channel

Delay and Sum beamformer which returns an enhanced signal obtained by delaying and summing the noisy input signals is performed by two scripts[1]:

- *nw_delaysum* is based on array processing written by Pirinen in 2004 and it was modified by Ferras in 2005. This script implements cross-channel time-delay estimation by using non-weighted cross-correlation.[1]
- *phat_delaysum* that is based on array processing and written by Pirinenin in 2004. It was modified by Ferras in 2005 to implement PHAT-weighted generalized cross-correlation (GCC-PHAT) for time-delay estimation.

[1]MATLAB source code for both Nw-DS and PHAT-DS is available on-line at http://www.icsi.berkeley.edu/Speech/papers/multimic/

Phase-error based filtering (PBF) is based on the time-frequency processing framework PBF performs speech enhancement in the short-time fourier transform (STFT) domain. Every input signal x(t) is decomposed into frames each of them which windowed and transformed into the frequency domain to be enhanced by means of a masking approach. After frequency-domain frame's processing, every frame is inverse transformed to the time-domain and added to the previously overlapped resynthesized frames [20]–[21]. Figure 4 shows a block diagram for a multi-microphone phase error based filtering.

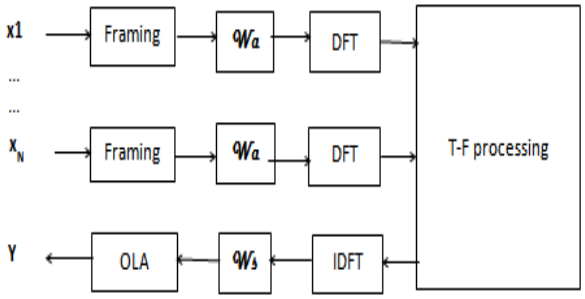

Figure 4.Multi-microphone phase-error based filtering diagram

For each input signal, its phase spectrum at frame m is calculated. First, for all possible pairs of frames phase-error is computed from their phase spectrums and used to modulate the amplitude spectrum. A masking function is then derived to weight the amplitude spectrum for each channel. Spectrums are later converted to Cartesian form and summed up shown the figure 5.

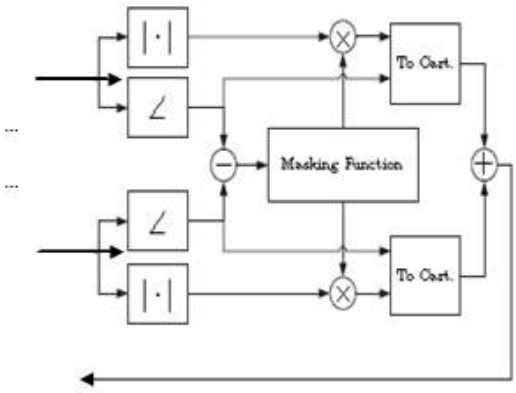

Figure 5. Masking approach block diagram

Phase-error based filtering which implemented by Marc Ferras in June 2005 is based on array processing .Cross-channel time-delay estimation

(TDE) is performed using PHAT-weighted generalized cross-correlation (due to its high sensitivity to time-alignment).

2.3 Our multi-microphone speech enhancement approach's model

So, referred to the section above our proposed approach can be divided into two models:

- Speech enhancement model via combination of preprocessing and multi-channel speech enhancement method: *Delay and Sum*as shown the figure 6 a).
- Speech enhancement model via combination of preprocessing and multi-channel speech enhancement method: *Phase-Error Based Filtering* as shown the figure 6 b).

a)

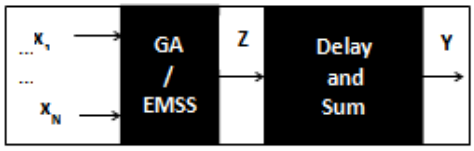

Figure 6.a) Model based on DS

b)

Figure 6.bModel based on PBF

3 Experiments and results

In this section we evaluate and compare our proposed approach for speech enhancement under a variety of real noisy environments.

3.1 Simulation conditions

To evaluate the performance of the proposed blind multi- microphone approach, described in section 2. We use some sentences taken from the Meeting Recorder Digits (MRD) database, this database consists of speech signals of a total of 576 sentences spoken by 144 speakers recorded in a real meeting room by four microphones [22].

The ICSI Meeting Recorder Digits Corpus provides a collection of connected digit speech data recorded in a real meeting room [23] in order to ease speech enhancement algorithm development in real-world environments. The core test set of the Meeting Recorder Digits (MRD) database (576 sentences) contains recordings of read connected digits performed by 144 speakers in a real meeting room with four microphones rated Channel 6, Channel 7, Channel E and F compressed as NIST SPHERE format.

In our simulations, all the speech signals are sampled at 16 kHz and were implemented using a frame length of 1024samples. To evaluate our approach, we selected 10 speakers noted mrd_10, mrd_11, mrd_12, mrd_13, mrd_16, mrd_20, mrd_30, mrd_137, mrd_140, mrd_144. Every record's speaker is performed by four microphones and mixed with 3 different types of background noise including white, car and babble noises from the NOIZEUS database (Hu and Loizou, 2007) at three different signal-to-noise ratios (SNRs) at four SNR levels (-5, 0, 5 and 10 dB). The noise babble and car are considered non-stationary.

The results obtained by our proposed approach are compared to multi-microphone speech enhancement methods without preprocessing: Delay-and-Sum and Phase-Error Based Filtering (2009). In addition we compare our approach's different model to find the best combination that can remove a maximum noise without speech distortion under a various criteria.

To test the performance of our multi-channel proposed speech enhancement approach**,** the enhanced speech signals are evaluated with an objective and subjective performance metrics under a various criteria.

3.2 Objective results

In 2006, Loizou and Yi Hu implement the composite objective measure proposed [24] to measure the performance of our multi-channel proposed speech enhancement approach. It gives three metrics: The predicted rating of overall quality (Covl), the rating of speech distortion (Csig) and the rating of background distortion (Cbak). The ratings are based on the 1-5 MOS scale. In addition, this function gives some objective speech quality measures: the Segmental SNR (SegSNR), the Log-Likelihood Ratio (LLR), the Perceptual Evaluation of Speech Quality (PESQ) and Weighted Spectral Slope (WSS).

The widely used objective speech quality measures: segmental SNR (segSNR), the perceptual evaluation of speech quality (PESQ) were evaluated in this study to assess the signals results' quality in two levels: channel's number, noise background's type [25].

Tables 1, 2, and 3 present respectively the SegSNR, PESQ, and COVL measures of the average results over number noisy channel criterion to evaluate our approach and to find the combination under each criterion.

The Segmental signal-to-noise (SegSNR) is a measure of signal quality objective. It is defined as the average of the signal-noise ratios calculated for plurality of segments.

$$SegSNR = \frac{10}{F}\sum_{k=1}^{F} log_{10}\left[\frac{\sum_{i=0}^{N-1} s^2(k,i)}{\sum_{i=0}^{N-1}[s(k,i)-\hat{s}(k,i)]^2}\right] (6)$$

where N is the length of each frame, and F is the number of frames $\hat{s}$ and x are the k^{th} frame of the denoising and original speech signal respectively.

Table 1. Average of SegSNR measures

	One ch. noisy	Two ch. Noisy	Three ch. Noisy	Four ch. noisy
nw_DS	-10,000	-9,989	-9,987	-10,000
Ga+nw_DS	-9,990	-9,970	-9,900	-8,645
MAP+nw_DS	-9,992	-9,980	-9,928	-8,604
MMSE+nw_DS	-9,997	-9,992	-9,985	-9,894
MMSE_S+nw_DS	-9,995	-9,988	-9,965	-9,614
SMPO+nw_DS	-9,995	-9,996	-9,958	-9,369
SMPR+nw_DS	-9,996	-9,991	-9,974	-9,731
phat_DS	-10,000	-9,989	-9,989	-9,999
Ga+phat_DS	-9,989	-9,968	-9,913	-9,380
MAP+phat_DS	-9,993	-9,979	-9,926	-9,013
MMSE+phat_DS	-9,996	-9,992	-9,985	-9,915
MMSE_S+phat_DS	-9,995	-9,988	-9,968	-9,733
SMPO+phat_DS	-9,995	-9,988	-9,956	-9,567
SMPR+phat_DS	-9,996	-9,990	-9,975	-9,803
PBF	-9,973	-9,986	-9,987	-9,984
Ga+PBF	-9,461	-9,522	-9,148	-8,668
MAP+PBF	-9,422	-9,440	-9,048	-8,575
MMSE+PBF	-9,761	-9,811	-9,748	-9,657
MMSE_S+PBF	-9,603	-9,696	-9,542	-9,352
SMPO+PBF	-9,579	-9,655	-9,444	-9,181
SMPR+PBF	-9,639	-9,733	-9,607	-9,473

The segmental signal-to-noise ratio (segSNR) reflects our approach's performance since under each criterion: one, two, three or four noisy channel each multi-channel speech enhancement method preceded by a pretreatment is better than without a pretreatment. According to segSNR results, the multi-channel phase-based filtering behaves better that multi-channel delay and sum by adding a pretreatment. Among all methods used as pretreatment, the estimator MAP outperforms all other methods with a large margin as a pretreatment since all multi-channel methods have their highest SegSNR scores.

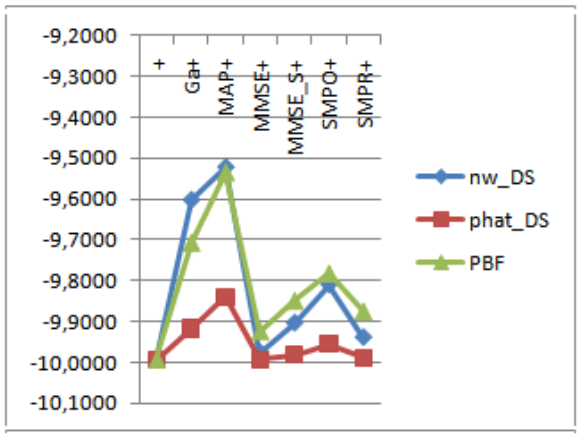

Figure 7. Average of SegSNR metric _white noise

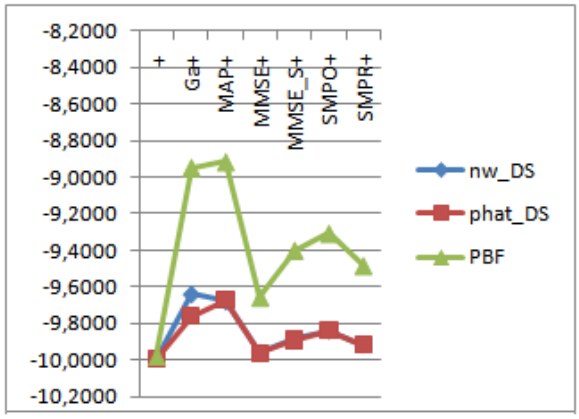

Figure 8. Average of SegSNR metric _non stationary noise

The two figures (7) and (8) confirm that our proposed approach based on PBF preceded by the estimator MAP is the most efficient method for enhancement of speech signals corrupted by both the additive white noise and the non-stationary noise [26, 27, 25].

In order to have a better idea about the estimated signal quality, the Perceptual Evaluation of Speech Quality (PESQ) metric has been used. The PESQ is defined in the ITU-T P.862 standard .The average Perceptual Evaluation of Speech Quality (PESQ) consists to map the estimated and the source signals onto an internal representation using a perceptual model. Also, the resulting of this metric measurement is equivalent to the subjective "Mean Opinion Score" (MOS) measured score.

It is considered as one of the reliable methods of objective test. It returns a score from 0.5 to 4.5. Table 2 illustrates the PESQ score obtained over four criteria: one, two, three and four noisy channel. Figure 9 and 10 gives the averaged PESQ scores for the above-mentioned methods, over all noise conditions.

Table 2. Average of PESQ measures

	One ch. Noisy	Two ch. noisy	Three ch. noisy	Four ch. noisy
nw_DS	2,472	2,407	2,374	2,134
Ga+nw_DS	2,700	2,576	2,462	2,149
MAP+nw_DS	2,671	2,535	2,403	2,123
MMSE+nw_DS	2,585	2,452	2,343	2,225
MMSE_S+nw_DS	2,622	2,489	2,374	2,180
SMPO+nw_DS	2,640	2,501	2,368	2,139
SMPR+nw_DS	2,609	2,474	2,352	2,191
phat_DS	2,480	2,425	2,162	2,162
Ga+phat_DS	2,718	2,564	2,102	2,100
MAP+phat_DS	2,684	2,515	2,098	2,098
MMSE+phat_DS	2,588	2,426	2,224	2,224
MMSE_S+phat_DS	2,633	2,460	2,164	2,164
SMPO+phat_DS	2,648	2,477	2,118	2,118
SMPR+phat_DS	2,616	2,447	2,183	2,183
PBF	2,359	2,237	2,122	2,082
Ga+PBF	2,406	2,261	2,120	2,094
MAP+PBF	2,388	2,207	2,166	2,397
MMSE+PBF	2,355	2,254	2,141	2,047
MMSE_S+PBF	2,371	2,239	2,094	1,986
SMPO+PBF	2,364	2,222	2,144	2,284
SMPR+PBF	2,360	2,242	2,114	2,031

Depending to PESQ results, we see that our proposed approach is still generally more effective in terms of perceptual quality. The table above shows that estimated signals obtained by using the proposed approach are better in term of the perceptual quality than the multi-channel methods mainly under the criterion one and two channel are noisy. So we can deduct that when we have a limit number of noisy channel, the best combination is the multi-channel Delay and sum preceded by the geometric approach for subtraction spectral in order to have an enhanced speech signals with high perceived quality.

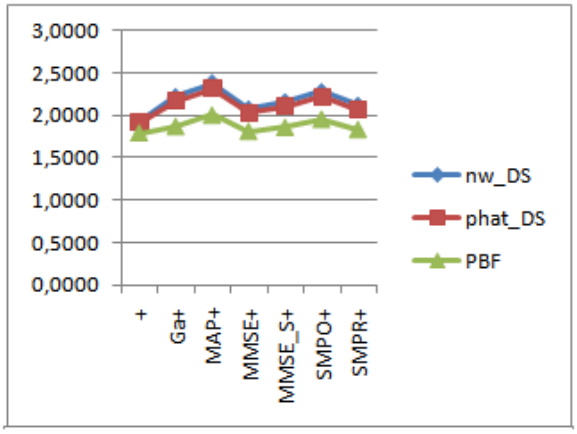

Figure 9. Average of PESQ metric _white noise

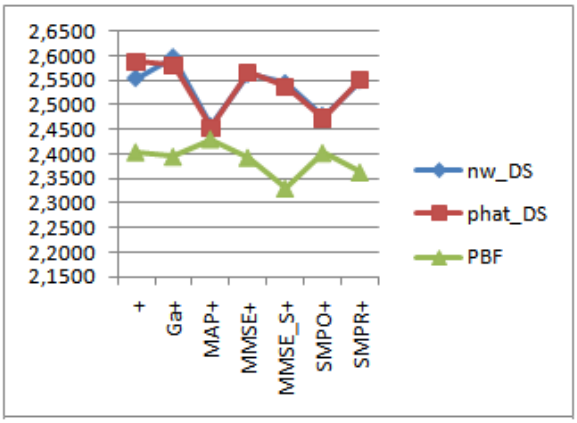

Figure 10. Average of PESQ metric _non stationary noise

The high PESQ scores perceived the quality of the enhanced speech. Our proposed approach based on Delay and sum *beamformer* is characterized by the highest PESQ scores for enhancement of speech signals:

- corrupted by stationary noise when this method is preceded by MAP
- corrupted by non stationary noise when this method is preceded by GA

Aiming to improve further our approach and to find the best combination for each criterion, we considered composite measures. Generally, composite objective indicators are obtained by linearly combining existing objective metric to form a new metric. The three new composite objective measures obtained are:

- Csig for signal distortion (SIG) formed by linearly combining the LLR, PESQ, and WSS measures
- Cbak for noise distortion (BAK) formed by linearly combining the segSNR, PESQ, and WSS measures

- Covl for overall quality (OVRL) formed by linearly combining the PESQ, LLR, and WSS measures

In this paper, we used the predicted rating of overall quality (Covl) to evaluate the quality estimated speech signals. This indicator is the result of combination of the evaluation measures in frequency domain, time domain and perceptual field. The ratings are based on the 1-5 MOS scale.

The Covl measure is described as follows:

$$Covl = 1.594 + 0.805 * PESQ - 0.512 * LLR - 0.007 * WSS \quad (7)$$

Note that, the objective speech quality measure: weighted spectral slope (WSS) is based on the difference between the adjacent spectral magnitudes in each frequency band. And, the objective speech quality measure: log-likelihood ratio (LLR) is determined by calculating the difference between the all-pole models of the enhanced and clean speech using the autocorrelation lags and LPC parameters. The results of the composite measure Covl are detailed in Table 3.

Table 3. Average of Covl measures

	One ch. noisy	Two ch. noisy	Three ch. noisy	Four ch. noisy
nw_DS	2,782	3,009	2,965	2,594
Ga+nw_DS	2,939	3,300	2,856	2,702
MAP+nw_DS	2,807	3,078	2,682	2,763
MMSE+nw_DS	2,471	3,187	2,979	2,726
MMSE_S+nw_DS	2,613	3,173	2,840	2,520
SMPO+nw_DS	2,685	3,099	2,640	2,554
SMPR+nw_DS	2,558	3,178	2,487	2,637
phat_DS	2,547	3,021	3,115	2,741
Ga+phat_DS	3,147	3,290	2,783	2,631
MAP+phat_DS	3,056	3,061	2,699	2,955
MMSE+phat_DS	2,734	3,169	3,190	3,065
MMSE_S+phat_DS	2,869	3,149	2,880	2,705
SMPO+phat_DS	2,934	3,078	2,576	2,850
SMPR+phat_DS	2,816	3,158	2,917	4,176
PBF	2.541	2.965	2,400	0,611
Ga+PBF	2.067	2.867	3,181	1,898
MAP+PBF	2.876	3.000	2,884	3,060
MMSE+PBF	2.187	2.987	2,683	1,420
MMSE_S+PBF	2.575	2.975	2,951	2,274
SMPO+PBF	2,703	2,800	2,817	2,963
SMPR+PBF	2.255	2.267	2,682	1,908

The proposed approach is characterized by the highest *Covl* values confirming its performance, as shown in table 3. These results confirm again the efficiency and consistency of adding a pretreatment to precede the multi-channel speech enhancement. As it can be seen in this table, the proposed approach phat_delay sum is characterized by the highest *Covl* scores showing that the enhanced speech by this method combined with a pretreatment has a better overall quality. As it is regarded the estimator SMPR followed by the multi-channel method phat_delay sum out performs all other methods with a large margin when the noisy channel's number is elevated.

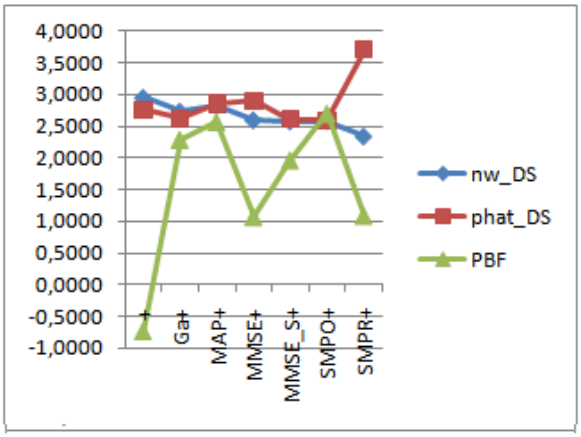

Figure 11. Average of Covl metric _white noise

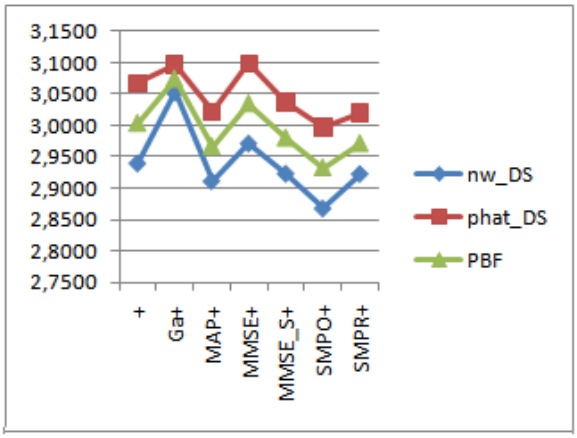

Figure 12. Average of Covl metric _ non stationary noise

The high Covl scores perceived the overall quality of the enhanced speech. This measure indicates the improvement of enhanced speech over the speech corrupted by stationary and non-stationary noise. The figure 11 shows that the best combination to remove white noise is the phat_DS preceded by the estimator SMPR. For the non_stationary noisy

speech, the enhanced speech that has the better overall quality is the result from the speech enhancement by the GA followed by the Phat_DS.

4 Conclusion

This work introduces a simple and efficient combination of mono and multi-channel speech enhancement approach. The method is based on adding a pre-processing technique precede a multi-channel speech enhancement method. We aim to apply a different class of mono-channel method in order to compare between them and to find the best combination that can remove a maximum noise without introducing artifacts. Our proposed approach is simple but gives better results to enhance speech corrupted by stationary and non-stationary noises compared to these multi-channel methods without pre-treatment. These multi-channel speech enhancement methods are evaluated using different objective measures like SNRSeg, PESQ and COVL in two levels: the noisy channel's number and the noise's type. From the results, it is evident that our approach is efficient for best noise removal and speech quality.

References:

[1] M.A. Ben Messaoud, A. Bouzid, and N. Ellouze, A New Biologically Inspired Fuzzy Expert System-Based Voiced/Unvoiced Decision Algorithm for Speech Enhancement. Cognitive computation. Vol. 8, No.1, pp.1-16, 2016.

[2] B.Xia and C.Bao , Wiener filtering based speech enhancement with Weighted Denoising Auto-encoder and noise classification, 2013.

[3] D.Cantzos, statistical enhancement methods for immersive audio environments and compressed audio, 2008.

[4] Kris Hermus, Patrick Wambacq, and Hugo Van hamme, A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition, KatholiekeUniversiteit Leuven, 3001 Leuven-Heverlee, Belgium ,received 24 October 2005; revised 7 March 2006; Accepted 30 April 2006.

[5] L. J. Griffith and C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, IEEE Trans. AntennasPropag, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[6] Y. Kaneda and J. Ohga, Adaptive microphone-array system for noise reduction, IEEETrands. Acoust. Speech, Signal Process., pp.2109–2112, 1986.

[7] H. F. Silverman and W. R. Pattterson, Visualizing the performance of large-aperture microphone arrays, in Proc. ICASSP'99, 1999, pp. 962–972.

[8] Saruwatari H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, Blind source separation based on a fast-convergence algorithm combining ICA and beamforming, IEEE Trans. Speech Audio Process., vol. 14, no. 2, pp. 666–678, Mar. 2006.

[9] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, Y. Ikeda, H. Hashimoto, and T. Morita, Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking, EURASIP J. Appl. Signal Process., vol. 2006, 2006, article ID 34 970.

[10] P. Comon, Independent component analysis, a new concept? Signal Process,vol. 36, pp. 287–314, 1994.

[11] Anuradha R. Fukane, Shashikant L. Sahare, Different Approaches of Spectral Subtraction method for Enhancing the Speech Signal in Noisy Environments,2011=11.

[12] Yang Lu, Philipos C. Loizou. A geometric approach to spectral subtraction, 2007. University of Texas-Dallas, Richardson, TX 75083-0688, United States Received 22 May 2007; received in revised form 18 January 2008; accepted 24 January 2008.

[13] Y. Ephraim and D. Malah.Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32, No. 6, pp. 1109-1121, 1984.

[14] Y. Ephraim and D. Malah."Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust, Speech, Signal Process, vol. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.

[15] Yang Lu, Philipos C. Loizou, Estimators of the Magnitude-Squared Spectrum and Methods for Incorporating SNR Uncertainty, 2011.

[16] Marc. Ferras.Font, Multi-Microphone Signal Processing For Automatics Speech Recognition in Meeting Rooms, 2005.

[17] A. Kareem, Z. Chao Zhu, Blind Source Separation Based of Brain Computer Interface System, 2014.

[18] Y.Zhang and W.H. Abdulla, A Comparative Study of Time-Delay Estimation Techniques Using Microphone Arrays,2005.The

University of Auckland, Private Bag 92019, Auckland, New Zealand.

[19] C. H. Knapp and C. Carter, The generalized correlation method for estimation of time delay, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 24, August 1976.

[20] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement", IEEE Transactions on Systems, Man and Cybernetics, vol. 34, August 2004.

[21] C. Y. Lai and P. Aarabi, "Multiple-microphone time-varying filters for robust speech recognition", Proc. ICASSP, 2004.

[22] A. Janin, J. Ang, S. Bhagat, R.Dhillon, J.Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, The icsi meeting corpus: Resources and research, NIST ICASSP, Meeting Recognition Workshop (Montreal, Canada), May 2004.

[23] The NIST Meeting Room Project, http://www.nist.gov/speech/test beds/mrproj/

[24] Philipos C. Loizou and Y.Hu, “Evaluation of objective measures for speech enhancement, Proceedings of INTERSPEECH-2006, Philadelphia, PA, September 2006.

[25] M.A. Ben Messaoud, et A. Bouzid, Speech Enhancement Based on Wavelet Transform and Improved Subspace Decomposition. Journal of Audio Engineering society (JAES). Vol. 63, No.12, pp.1-11, 2015.

[26] Sandhya Hawaldar and Manasi Dixit. Speech Enhancement for Nonstationary Noise Environments. Signal & Image Processing: An International Journal, Vol. 2, No.4, 2011.

[27] Bolimera Ravi and T. Kishore Kumar. Speech Enhancement Using Kernel and Normalized Kernel Affine Projection Algorithm. Signal & Image Processing: An International Journal, Vol. 4, No.3, 2013.

[28] M.Ravichandra Kumar and B.RaviTeja. A Novel Uncertainty Parameter SR (Signal to Residual Spectrum Ratio) Evaluation Approach for Speech Enhancement. Signal & Image Processing: An International Journal, Vol. 4, No.3, 2013.