International Journal of Computer Applications (0975 - 8887)
Volume 179 - No.17, February 2018

A Survey on Techniques for Enhancing Speech

Tayseer M. F. Taha
Colleage of Computer Sciences
Sudan University for Sciences and Technology, Khartoum, Sudan

Ahsan Adeel
Dept. Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling
Stirling, Scotland, UK

Amir Hussain
Dept Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling
Stirling, Scotland, UK

ABSTRACT

Speech enhancement is used in almost all the modern communication systems. It is obvious that when speech is being transmitted, its quality may degrade due to interference in the environment it is passing through. Some of the interferences that may affect the speech quality of transit include acoustic additive noise, acoustic reverberation or white Gaussian noise. This paper focuses on the techniques that appeared in the literature to enhance the signal of speech. Various methods used include wiener filter, statistical methods, subspace method, basic spectral subtraction method and spectral subtraction. In this paper authors will discuss various such methods along with their advantages and disadvantages. The discussion will also review the studies conducted by other researchers on other machine learning techniques, such as Neural network, Deep Neural Network ,Convolution Neural Networks and optimization techniques which used for the enhancement of speech.

General Terms

Signal processing, Machine learning, optimization

Keywords

Conventional speech enhancement methods, Adaptive filtering methods, Multi-modal methods

1. INTRODUCTION

Speech enhancement is a vital element in the communication equipment. It refines speech and reduces noise, it is used in a variety of domains for example to assist in hearing and other applications such as mobile phones, teleconferencing system, hearing aids, voice communication systems.

Speech enhancement is closely related to speech restoration because it reconstructs and restores the signal after degradation[1]. However, there is a slight difference between the speech restoration and speech enhancement. Restoration of speech means to convert the noisy signal back to its original form- prior to noise addition. Speech enhancement on the other hand helps in refining the original signal to be better. Also, an original undergirded speech signal cannot be restored but can be enhanced [2]. The aim of these speech enhancement algorithms, to improve perceptual aspects of the speech signal, that is degraded by the additive noise such as overall quality or intelligibility with the aim of reducing listener fatigue [3] [4].

Enhancement of speech can be used in different settings, such as in areas where there is an interfering background noise in a building, in noisy streets or roads where there are motor vehicles passing. These interference noises degrade the quality original speech in such a way it does not remain clear anymore. An important context that needs to be addressed for speech enhancement includes the compression of speech bandwidth systems[5], this is mostly used in the decoding of digital channels of communication. This technique is also needed for the decoding of the speech, which includes integration of data and voice networks, including speech bandwidth compression systems that plays an important role in speech communication systems.

Ravi and Subbaiah [2]conducted a survey on single channel speech enhancement methodologies,[6] considered single and multi-channel speech enhancement in their review paper. Amole and Dhonde [7] presented a review on the spectral subtraction method and its modification.The Authors in [8] addressed time and transform domain speech enhancement methods .Statistical based techniques for speech enhancement reviewed by sunnydayal and Sivaprasad in [9].

To the best of authors knowledge there is no previous work in the literature which categories the speech enhancement in this way. In this study authors classify speech enhancement methods into four categories:Conventional methods, Adaptive filtering methods, Machine learning methods (this includes Adaptive filtering using optimization techniques), and Multi-modal methods.

This paper is organized as follows:Section 1 gives an introduction to the problem and a general overview. Section 2 gives the reader an understanding of the types of audio enhancement categories. Section 3 describes the kind of noise considered in the research. Sections 4 - 7 discuss four basic approaches of speech enhancement, the conclusion will come in section 8 .

2. CATEGORIES OF AUDIO ENHANCEMENT METHODS

According to Lim JS and Oppenheim AV [10] and Loizou [3], classification of speech enhancement techniques can be based on the following:

(1) Type of the algorithm used, which can either be adaptive or non-adaptive,

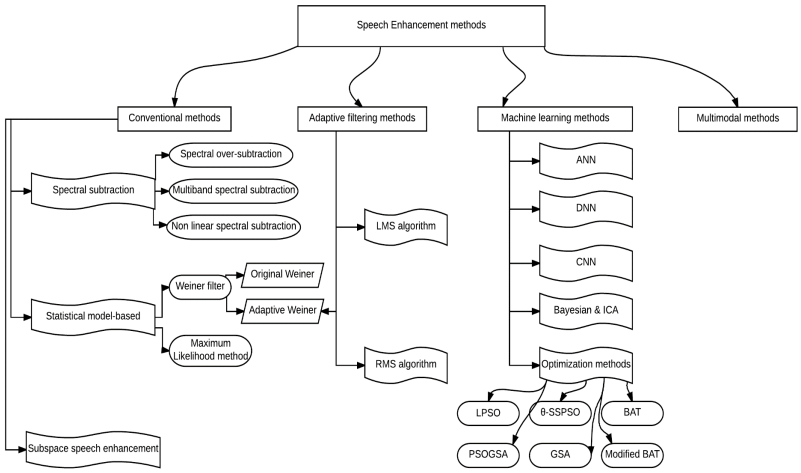

Fig. 1. Popular techniques reviewed in the current study

(2) The input channels involved, which can either be single, dual or multiple
(3) Whether it is uni-modal or multimodal

The following parts of this section, state the differences between the above mentioned techniques.

2.1 Difference between adaptive and non-adaptive speech enhancement

If additive noise is present in a speech signal, then common practice is to pass it through a filter that removes the noise while minimally interfering with the original signal component. This is called direct filtering. Initial work in this domain of optimal filtering, was done by Weiner [11] and was extended and enhanced by [12] and others. Filters used for direct filtering can be either Fixed or Adaptive.

—Fixed filters to design these, it is important to have prior knowledge of both the signal and the noise. It passes frequencies present in the signal and discards the frequency band occupied by the noise.

—Adaptive filters : can adjust their impulse response to filter out the correlated signal component in the speech input. They require almost no prior knowledge of the characteristics of signal and noise. (in case, the signal is narrowband and noise is broadband- or vice versa- no prior information is needed; otherwise they require to know desired response of the signal). They can adaptively track the signal in the presence of non-stationary conditions.

2.2 Difference between monaural and binaural speech enhancement

Single channel enhancement, also known as Monaural enhancement, is for situations where only one input channel is present such as mobile telephony [2]. In multichannel speech enhancement, the noisy observations are obtained from two or more sensors. If there are only 2 channels in multichannel system, then it is also called binaural enhancement. It has two types:

(1) Supervised methods (like NMF, HMM) where noise and speech are modeled according to training samples [13].
(2) Unsupervised methods (like transform domain approaches/Wiener filter/Kalman filter) where no training samples are needed. Neither is needed any prior information about the signal or the calculation of Noise Power Spectral Density [14][15].

Figure 2 below shows the types of single channel enhancement methods[2].
Multi channel algorithms show better performance with respect to substantial speech reception threshold scores when the target signal and the noise source are separated [16]. However, in practical scenarios, these requirements might not always be fulfilled, and single channel algorithms are preferred for devices, such as hearing aids in which the number of microphones is usually limited to two and the two microphones are on the same side of the head (thus recording the same signal) [67].

2.3 Difference between uni-modal and multi-modal speech enhancement

According to Monaci Gianluca[18], use of internal stimuli in senses enables individuals to identify different perceptions in the environments that they live in. Humans integrate acoustic and visual signals[19] [20][21] [22] or tactile and visual inputs [23] [24].
If audio perception is enhanced using just the auditory sense, then this can be referred to as unimodal audio enhancement. On the other hand, when audio perception is enhanced by a couple of other senses such as the auditory sense and the visual sense, then it is referred to as multimodal speech enhancement.
Multimodal speech enhancement on the other hand is where audio signal is enhanced by senses other than the auditory sense- for example speech/ vision/ language/ text. Hence it significantly enhances performance [25].
In this review paper, authors are going to survey all the prominent work that has been done in the domain of speech enhancement up till now.

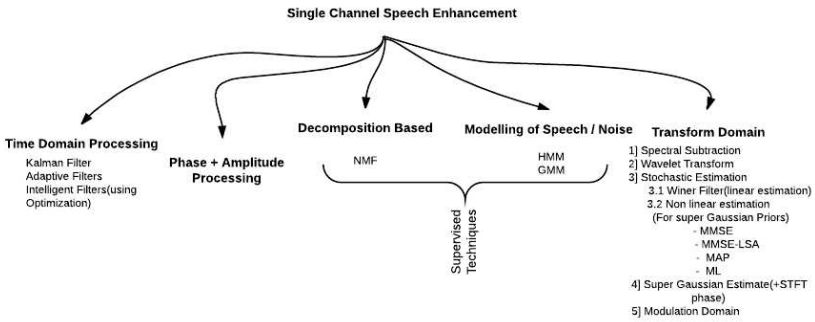

Fig. 2. Popular techniques reviewed in the current study

3. NATURE OF NOISE CONSIDERED IN CURRENT RESEARCH

When processing a speech signal, we may come across a number of types of noise that it may be contaminated with. Common types of noise that can be added to speech signals are enlisted below by Lakshmikanth.S in [57]:

(1) Background noise: environmental distortion or noise of cars on road for example.
(2) Echo: that occurs in closed spaces with bad acoustics
(3) Acoustic echo: also known as audio feedback: it sometimes occurs in two-way communication when the telephone microphone captures the speech of the person on the other side of the telephone.
(4) Amplifier noise: if amplifier produces even a little additional thermal noise, it becomes hugely noticeable after amplification process. Such noise is called amplifier noise.
(5) Quantization noise: it is created as part of the transformation process of the signal from analogue to digital domain, interference occurs in sampling while rounding up real values of analogue signal.
(6) Loss of signal quality: caused by coding and speech compression.

Because of the huge amount of works reported in this field, this survey will only consider the case when the noise is additive and independent of the clean speech. Survey of techniques purely for echo cancellation source separation case studies are not reviewed in present research. Various speech enhancement techniques have been put forth for the purpose of improving perceptual aspects of a speech signal, that has been degraded by additive noise. These techniques improve overall quality and intelligibility, and reduce listener fatigue [3][4]

4. CONVENTIONAL METHODS OF SPEECH ENHANCEMENT

This section will discuss different single channel speech enhancement methods.

4.1 Spectral Subtraction Method (single channel speech enhancement)

Spectral subtraction method is one of the oldest methods of single channel speech enhancement. It is considered to be among the first algorithms in this domain[15]. It is simple and effective in elimination of stationary background noise.Its limitation is that it suffers from narrow-band tonal- commonly called 'musical noise' [27]. Various modifications of spectral subtraction have been proposed to improve its results [51]. If clean speech signal $x(n)$ additive noise signal which is uncorrelated with the clean speech $d(n)$ then the signal corrupted by the noise $y(n)$ can be written as:

$$y(n) = x(n) + d(n) \tag{1}$$

Since speech signal is non-stationary, so noise component is processed frame-by-frame in the frequency domain [28]. Discrete time Fourier transform for both sides yields:

$$Y(\omega) = X(\omega) + D(\omega) \tag{2}$$

To get spectrum of enhanced speech, the method by [3] is used:

$$|\hat{X}(\omega)| = |Y(\omega)| - |\hat{D}(\omega)| \tag{3}$$

Where $|\hat{X}(\omega)|$ is the estimated speech short time magnitude, $|Y(\omega)|$ is the noisy speech short time magnitude and $|\hat{D(\omega)}|$ is a noise spectral magnitude estimate computed during non-speech activity. The power spectrum subtraction is then given by:

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{D(\omega)}|^2 & if |Y(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0 & otherwise \end{cases} \tag{4}$$

Much work has been done to suppress noise that occurs as a side product to spectral subtraction method using varied forms of spectral subtraction: :spectral over-subtraction [14], multi-band spectral subtraction [29] non- linear spectral subtraction [30], iterative method[31] spectral subtraction based on perceptual properties[32].The following subsections explain these variants of spectral subtraction methods.

4.1.1 Spectral over-subtraction method. A further modification in basic spectral subtraction method [15] resulted in a variation commonly known as Spectral Over-Subtraction Method [14]. Following parameters were introduced to reduce noise:

(1) Over- subtraction factor: which has control over the amount of noise power spectrum subtracted from the noisy speech power spectrum.
(2) Noise spectral floor: which restricts the resultant spectral component from increasing above a preset minimum spectral flow value [3].

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\hat{D}(\omega)|^2 & if |Y(\omega)|^2 > (\alpha+\beta)|\hat{D}(\omega)|^2 \\ \beta|\hat{D}(\omega)|^2 & otherwise \end{cases} \quad (5)$$

where $\alpha \geq 1$ and $0 \geq \beta \geq 1$.

4.1.2 Multi-band spectral subtraction. The Multi-band spectral subtraction is another variation of this type, where the speech spectrum is partitioned into several non-overlapping regions, and then spectral subtraction is applied on each band separately. Clean speech spectrum is represented by the following math model[3][29]:

$$|\hat{X}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i\delta_i|\hat{D}_i(\omega)|^2 & if\ |Y_i(\omega)|^2 > 0, \\ k_i < \omega < k_{i+1} & \\ \beta|\hat{Y}_i(\omega)|^2 & otherwise \end{cases} \quad (6)$$

where k_i, and k_{i+1} are the beginning and the ending of the frequency bins of the i^{th} frequency band, α_i the over subtraction factor of the i^{th} band, and δ_i is a tweaking factor in the i^{th} band.
Band -specific over-subtraction is represented as a function of segmented SNR_i of the i^{th} frequency band. Following is its mathematical representation[3]:

$$\alpha_i = \begin{cases} 5 & if\ SNR \leq -5 \\ 4 - \frac{3}{20}SSN_i & if\ -5 \leq SNR_i \leq 20 \\ 1 & if\ SNR > 20 \end{cases} \quad (7)$$

and δ_i the control of each band is calculated as:

$$\delta_i = \begin{cases} 1 & if\ f_i \leq 1kHz \\ 2.5 & if\ 1kHz \leq f_i \leq \frac{f_s}{2} - 2kHz \\ 1.5 & if\ f_i > \frac{f_s}{2} - 2kHz \end{cases} \quad (8)$$

where f_i is the upper frequency of the i^{th} band, and f_s is the sampling frequency[29, 3].

4.1.3 Non-linear spectral subtraction. Lockwood and Boudy [30] introduced a modified version of over-subtraction by proposing a technique where the nature of subtraction process is nonlinear and the over-subtraction factor frequency depends upon frame SNR [3].
Over the years, many modifications have been suggested to vary the original method of spectral subtraction algorithm in order to reduce the musical noise that occurs. Hu et. al [33] proposed the combination of comb filtering and spectral smoothing along with formant intensification for the enhancement of noisy speech. This brings significant improvements over the classical method in terms of perceived sound quality. A limitation was that these researchers only performed testing on two noise types (white Gaussian and car noise). A major drawback was that the and the simulation analysis showed contradiction between objective and subjective measures.
A further modified multi-band spectral subtraction (I-MBSS) algorithm was proposed in [34] for the enhancement of audio speech signal in different noise conditions. Here, noise-speech spectrum is partitioned into K non-overlapping bands and spectral over-subtraction method was applied independently on each band. Experimental analysis was done on different types of added noise. Dataset for simulations was NOIZEUS speech corpus. However, the simulations were not conducted on extremely low and high SNR levels.
A new algorithm comprising Generalized Sidelobe Cancellation (GSC) combined with spectral subtraction speech enhancement was put forth by Yu et al In [35]. Their research showed that the output signal from GSC module removes the remaining non-coherent noise upon filtering. They selected the additive noise from NOISEX-92, and their method showed prominent improvement in speech quality. The method was feasible enough to yield stable results. Cao et. al [36] designed a modulated filterbank that was oversampled to divide the time series into equal sub spaced bands. The authors in [37] made use of a weighted recursive averaging method to approximate the noise power spectrum, after which a multiband subtraction was applied on noise that was added to speech signal. An auditory masking threshold was computed with the estimated speech signal. In this way, the subsequent associated subtraction factor was adjusted. Experimentation proved their algorithm to be effective enough to enhance a signal that had been corrupted by white noise and also by musical noise. Only drawback was that they did not perform testing with multiple objective measures. They used only Itakura-Saito Distance (IS) objective measure to evaluate the proposed method. Another non-linear spectral subtraction technique for speech enhancement was used by Prabhakaran et.al [38]. Three types of noise were used to evaluate the proposed approach (pink noise, white noise, and Volvo noise). These samples were taken from dataset of TIMIT & NOIZEUS corpus. Islam et.al[39] did research on a speech enhancement approach that was formulated on modified spectral subtraction process carried out on time magnitude spectral. Extensive testing was done on NOIZEUS database. Simulation results showed that their proposed method is only suitable for higher segmental SNR. Bharti et. al[40] presented an adaptive method of noise cancellation and signal estimation that is based on short term energy. In this technique, noise spectrum is continuously updated. NOISEUS speech corpus was used for the evaluation of the proposed approach. This method works well for stationary and also for non-stationary noise. Only drawback is that system performs is not good at 0 db stationary noise.

4.2 Statistical model-based algorithms

statistical model-based methods considered as one of the common techniques for speech denoising. The method from this type operates in the noisy domain. In this method, noise is reduced by modifying the frequency spectrum of the noise signal[41]. The two algorithms of this category are:

(1) The Wiener algorithms.

(2) Minimum mean square algorithms.

4.2.1 Wiener filter speech enhancement . Weiner filter operates in the frequency domain. Its modified version, called adaptive Weiner filter operates in the time domain. The Original Weiner filter, Wiener[11] introduced wiener filter in 1949. It is quite similar in nature to the spectral subtraction method. It trades the subtraction step of spectral subtraction with an approximation of the signal spectrum of clean signal with a minimum mean square error (MMSE). It also involves the computation of short -time Fourier transform (STFT). The technique minimizes the MSE between the approximated signal magnitude spectrum $\hat{D(\omega)}$ and the original signal magnitude spectrum $D(\omega)$.

The optimal wiener filter is represented by [**?**]:

$$H(\omega) = \frac{D_s(\omega)}{D_s(\omega) + D_n(\omega)} \tag{9}$$

where $D_s(\omega)$ and $D_n(\omega)$ is the estimated power spectra of the noise-free signal and the background noise(noise assumed to be uncorrelated and stationary). Finally, the speech is enhanced by:

$$\hat{D}(\omega) = X(\omega)H(\omega) \tag{10}$$

Almajai and Milner [42] proposed visually derived wiener filter for speech enhancement, which exploits the audiovisual correlation. Weiner filters have the additional point that they can be used for both single channel and dual/ multiple channel. Jeub and Vary[43]introduced a novel speech enhancement algorithm for the binaural de-reverberation. The approach was solely based on a multiplex Wiener filter, which is augmented in order to be implemented to the digital hearing aids and binaural telephony headphones. It is further divided into two variations:

(1) The first, is an enhanced logical model which is made by taking the observational effects of the head in consideration

(2) Second, is a structure of binaural input-output which does not affect the key binaural signals and henceforth, also has localization ability. The Assessments done with the measured binaural room impulse responses indicate that this approach is very proficient in reducing reverberation.

Table*4.2.1* shows the pros and cons of the wiener filter.
Adaptive Weiner filter is dependent on the variation of the filter transfer function from sample to sample according to speech signal statistics (mean/variance). It was proposed by Abd El-Fattah et .al[44][45], and It works in time domain instead of the frequency domain (original wiener filter works in frequency domain). A recursive noise estimation approach is used for noise estimation. Sulong et al.[46] combined the process of the compressing sensing method and wiener filter for the noise reduction.

4.2.2 Maximum Likelihood method. This method was brought forth by Ephraim and Malah[47][49]. Yariv Ephraim [48] proposed that based on the estimation of the short-time spectral amplitude (STSA). The author derived the MMSE STSA estimator, based on modeling noise and speech spectral components as statistically independent Gaussian random variables and analyzed the performance of the proposed STSA estimator and compared it with Wiener estimator based STSA estimator.
MMSE STSA estimator is used to examine signals based on the quality or strengths in the deafening conditions and in the areas where there is the uncertainty of the presence of the signals. To construct the enhanced signal, an MMSE STSA estimator is used with the compound exponential of the deafening segment. Apriori probability distribution of the speech and noise Fourier expansion coefficients should be known to derive MMSE STSA estimator. The same authors Ephraim and Malah [47]also proposed the short-time spectral amplitude (STSA) estimator for speech signals that MMSE of the log spectra and inspect it in enhancing noisy speech. The results evaluated that the new estimator is good in improving noisy speech. The main Ephraim and Malah noise suppression rule is expressed in the following part. Neglecting the time and the frequency indexes(l, ω) for limitation of the notation, the suppression value $G(l, \omega)$ applied to each short-time spectrum value $X(l, \omega)$ to give[48]:

$$G(l,\omega) = \frac{\pi}{2}\sqrt{\Big(\frac{1}{1+R_{post}}\Big)\Big(\frac{R_{prio}}{1+R_{prio}}\Big)} \tag{11}$$
$$*M\Big[(1+R_{post})\Big(\frac{R_{prio}}{1+R_{prio}}\Big)\Big]$$

where M is a function based on the modified Bessel functions of zero and first order.

$$M[\theta] = \exp(-\frac{\theta}{2})\Big[(1+\theta)I_0\Big(\frac{\theta}{2}\Big) + \theta I_1\Big(\frac{\theta}{2}\Big)\Big] \tag{12}$$

The formulations of the a-priori SNR $(R_{prio}(l, \omega))$ and a-posteriori SNR $(R_{post}(l, \omega))$ respectively (for each value of the time and frequency indexes) are given below:

$$R_{post}(l,\omega) = \frac{|X(l,\omega)|^2}{D(\omega)} - 1 \tag{13}$$

$$R_{prio}(l,\omega) = (1-\alpha)P[R_{post}(l,\omega)] + \alpha\frac{|G(l-1,\omega)X(l-1,\omega)|^2}{D(\omega)} \tag{14}$$

where $D(\omega)$ is the noise power at frequency ω, with $P[x] = x$ if $x \geq 0$ and $P[x] = 0$ otherwise. $(R_{prio}(l, \omega))$ is an estimate of the SNR that takes into account the current short-term frame with weight $(1 - \alpha)$ and the noise reduced previous frame with weight α[49].

4.3 Subspace speech enhancement methods

Another type of speech enhancement methods is when speech estimation is considered as a constrained optimization problem. This approach was introduced by Ephraim and Van [50], and by Loizou in [51], where the noisy speech signal vector cosmos is decayed into two subspaces i.e. a signal subspace and a noise subspace. The Singular Value Decomposition (SVD) or the eigenvalue decomposition (EVD) is used to decompose the noisy signal into a noise signal and a speech signal. Surendran et.al [52] proposed a signal subspace speech improvement algorithm using the perceptual feature by using the frequency disguising property of human auditory system of frequency masking property of human auditory system[53]. A cue to spectral deviation ratio (SSDR) standardization is used for the reduction of the spectral misrepresentation. Samples of speech are used from the NOIZEUS database for the assessment of the introduced algorithm. The results of their experiments showed the effectiveness of their algorithm in speech enhancement compared to some benchmarks speech enhancement methods. An approach of the subspace method on the basis of Karhunen-Love transform and customs principal component analysis was proposed by Wang et. al[54] for the reduction of noise in different noisy environments. They used objective assessment measures (including Segmental SNR (SegSNR), Weighted Spectral Slope (WSS), the Log-Likelihood Ratio (LLR), Log Spectral Distance (LSD) and Perceptual Evaluation of Speech Quality (PESQ) to assess the performance of their algorithms. It was shown that their algorithm was more operative for white noise than colored noise. Performance was not good for SNR greater than 10dB. An effort was made by Sun et al [55] to introduce an algorithm based on joint low-rank and sparse matrix decomposition (JLSMD). It is different from the preceding subspace algorithms in its decomposition nature. Results showed

Table 1. Pros and cons of Wiener filters algorithms

Pros	Cons
• The algorithm safeguards a dereverberation performance that does not depend on the azimuth angle of the speech source • It preserves binaural cues • This algorithm is proficient enough to significantly reduce the effects of the reverberation especially in the rooms that are highly reverberant. • The algorithm is less complex in terms of computing calculations	There is a wide room for improvement in its performance in rooms with moderate reverberation

that their algorithm is better in improving the overall quality of the enhanced speech, however, noise reduction still had room for improvement.
Table 4.3 summarizes the advantage and disadvantages of the main convention speech enhancement methods[56]:

5. ADAPTIVE NOISE CANCELER(ANC)

Basically, an ANC denotes the electromechanical or electro-acoustic procedure of abandoning acoustic disruption to produce a softer environment Lakshmi et. al[57]. ANCs create and use an 'anti-noise' signal with same amplitude and opposite phase. The Adaptive Noise Canceler has been used in a number of applications such as hearing protectors, headsets, etc. ANC can be globalized to a multichannel system, which can be seen as a generalized beamforming system. An adaptive noise canceler was initially introduced by Widrow and Stearns [58]. It requires minimum two microphones founded on the basis of the obtainability of orientation channel(s) which are features of associated samples or references of the polluted noise. An estimate of the noise is produced with the help of adaptive filter by utilizing the reference microphone output. Its output is then deducted from the primary microphone output (signal + noise). The output of the canceler is used to regulate the tap weights in the adaptive filter. With the help of an adaptation algorithm, ANC minimizes the mean square error value of the output. It generates output which is the best approximation of the anticipated signal in the sense of minimum mean square error.
Adaptive filters fine-tune their coefficients to diminish an error signal and can be grasped as finite impulse response (FIR), infinite impulse response (IIR), lattice and transform domain filters. Least mean square [12] is the most common adaptive algorithm. Most sound foundations tend to be broadband in nature and while a huge share of the energy is focused in the lower frequencies, they also tend to have noteworthy high frequency components. Pros and Cons of this method are recruited below[57].

5.1 Types of Adaptive Noise Cancellation filters

An adaptive filter is a device which is used for computational purpose and it endeavors to create and establish the association between two signals in real time in an iterative style. An adaptive filter is defined by following phases [78]:

(1) The signal being treated by the filter;
(2) The configuration that describes how the output signal of the filter is calculated from its input signal.
(3) The limitations within this structure that can be iteratively altered to change the filters input-output association
(4) The adaptive algorithm that defines how the limitations are attuned from one time prompt to the subsequent.

5.1.1 Least Mean Squares (LMS) Algorithm. One of the extensively used techniques for the adaptive filtering is the LMS algorithm. Its foundation is credited to Windrow and Hoff [58] and Haykin [60]. It is based on the approximation of the gradient in the direction of the optimal solution using the arithmetical properties of the input signal. A noteworthy feature of the LMS algorithm is its straightforwardness. In this algorithm filter weights are rationalized with each new sample as required to meet the anticipated output. An acoustic echo canceler (AEC) is used to remove acoustic response from the loudspeaker to the microphone in the applications such as hands-free telephony, tele-classing and video-conferencing.
Adaptive filters with thousands of coefficients are used for room acoustic echo cancellation. Transform domain adaptive filter results in a noteworthy decrease in the computational weight. In [61], authors present Hirschman Optimal Transform (HOT) based adaptive filter for elimination of echo from audio signals. In order to test the efficacy of the proposed method, adaptive algorithms based on LMS, Normalised least mean squares(NLMS), Discrete Fourier Transform(DFT)-LMS and HOT-LMS were implemented and tested in this echo cancellation application. Their experiments proved that HOT based LMS adaptive filter is computationally effective and has fast convergence as compared to LMS, NLMS and DFT-LMS. For the cancellation or suppression of the assorted noise, they used this spectrogram technique to sense and eradicate noise. Through the method described in this paper, 12dB or more SNR can be attained, and noise reduction coefficient becomes greater than 0.9.

5.1.2 Recursive least squares (RLS) Algorithm. As, the adaptive filter is based on the alteration of the treated signal, it uses an adaptive algorithm for the alteration of the filter limitations and structure [62]. Normally, just the filter coefficients are altered and the remainder of the filter structure is same.
RLS adaptive algorithm for noise cancellation uses the error signal to regulate the weight coefficients of the adaptive filter, and therefore attains a filter output that is an estimate of the interference signal, and then uses the mixed signal with the noise component to subtract the filter output in order to acquire the strong signal and achieve the output of eliminating the noise signal.
This method was used by [62] to carry out research on speech enhancement using signals which had periodic noise mixed with impulse noise. Time-frequency spectrogram was used by them in order to pre-process the noisy signal, then they passed the signal through RLS adaptive noise reduction system to terminate the noisy component. The authors in[63] achieved the same results, this

Table 2. Advantages and disadvantages of main convention speech enhancement methods

Sppech enhancement	Advantages	Disadvantages
Spectral Subtraction	The Spectral Subtraction is effective in computational and has modest contrivance to control trade-off between speech misrepresentation and remaining noise	The introduced musical noise is disadvantage
MSSE estimator	It has less computational assets and resources	There is the absence of the mechanism in order to control trade-off between speech distortion and remaining noise
Wiener Filter	Reasonable computation load	There is the absence of the mechanism in order to control trade-off between speech distortion and Remaining noise
Subspace	It delivers a mechanism to control trade-off between speech distortion and remaining noise	It results in the heavy computational loads

Table 3. Pros and cons of Adaptive noise canceler

Pros	Cons
The customary wideband algorithms of ANC produce the best results in the lower frequency bands	• As the bandwidth and the center frequency of the noise upsurges, their performance depreciates quickly. • The algorithms are not appropriate for the multimodal error surface, and they provide a single likely solution for each reiteration according to the generated error. • It is necessary to have a frequency dependent noise cancellation system to avoid adversely affecting the desired signal in order to combine the ANC system with other communication and sound systems

approved again that RMS is superior compared to NLMS for noise cancellation.
A new dual forward blind source separation (FBSS) algorithm was introduced by [64] which was based on the use of the recursive least square algorithm to update the cross-filters of the forward structure. This algorithm combines the good features of both-FBSS and the RLS algorithm. This DFRLS algorithm was used by them in speech enhancement and acoustic noise reduction application. Their method showed good results as compared to dual forward normalized least mean square (DFNLMS) algorithm with respect to segmental signal to noise ratio (SegSNR), the cepstral distance (CD), the system mismatch (SM) and the segmental mean square error (SegMSE). A summery is given in table *5.1.2* for the advantages and dis-advantages of both the Least mean square(LMS) and the Recursive least square(RLS)

6. MACHINE LEARNING APPROACHES TO SPEECH ENHANCEMENT

6.1 Neural Networks for speech enhancement

A speech enhancement algorithm was evaluated by (Goehringa, et al [67]. It was based on neural networks speech enhancement (NNSE) to improve speech intelligibility in noise for cochlear implant (CI) users. The algorithm decays the noisy speech signal into time-frequency divisions, extracts a set of auditory characteristics and inserts them to the neural network to yield an approximation of frequency channels that contain more perceptually significant statistics (higher signal-to-noise ratio). This approximation is used to reduce noise-dominated component and retain speech-dominated components for electrical stimulation. The architecture and low processing delay of the NNSE algorithm make it appropriate for application in hearing devices.

6.2 Deep Neural Networks for speech enhancement

A regression-based speech enhancement framework was presented by [68]. It used deep neural networks (DNNs) with a deep architecture having multiple-layers. A Restricted Boltzmann Machine pre-training scheme was introduced to prepare the DNN. A huge training set is fundamental to learn the rich structure of the DNN. Using more acoustic framework statistics is shown to improve performance and make the enhanced speech less intermittent. Multi-condition training can deal with speech augmentation of new speakers, hidden noise types, numerous SNR levels under different noise circumstances, and even cross-language generalization. Compared with the SNN-based and Log-MMSE methods, noteworthy enhancements were attained on the TIMIT corpus. On average, 76.35% subjective preference was attained due to the nonappearance of musical noise in improved speech. Subsequently, the same authors introduced an altered version of this work in[69]. This was an administered technique to improve speech by means of finding a mapping function between noisy and clean speech signals based on deep neural networks (DNNs). This method can well suppress extremely non-stationary noise, which is hard to handle in general. Additionally, the subsequent DNN model, trained with synthetically created data, is also effective in dealing with noisy speech data logged in real-world situations without the generation of the infuriating musical artifact usually seen in conventional enhancement methods. Multi-condition training with many kinds of noise categories can attain a good generalization proficiency to hidden noise surroundings. By doing so, the proposed DNN framework is also influential in managing the non-stationary noises in real-world situations. Compared with the Log-MMSE technique, noteworthy enhancements were attained across different hidden

Table 4. Advantages and dis-advantages of LMS and RLS

	Pros	Cons
LMS	• The implementation of the LMS algorithm is simple to use and easy[61]. • HOT based LMS adaptive filter is computationally effective and has fast convergence as compared to LMS, NLMS and DFT-LMS[61]	• Simple LMS has sluggish convergence and gradient noise amplification[65]
RLS	• SNR can increase up to 10dB or more noise • reduction coefficient can reach more than 0.9 [62]. • Good noise reduction can be achieved.[64] • RLS has quicker rate of convergence as compared to LMS[66] • It has reduced steady-state error[66] • Its spectral characteristics are enhanced better than those of LMS[63]	• Its effects are mostly restricted to the periodic noise, the low-frequency noise signal [62]

noise situations. The sole disadvantage was that training data was too limited to cover a wide range of various acoustic scenarios, such as speaker and language inconsistencies.

6.3 Convolution Neural Networks for speech enhancement

In 2016, a model based on signal-to-noise ratio (SNR) aware Convolution Neural Network (CNN) was put forth for Speech Enhancement (SE) [70]. This CNN model can efficiently handle the local temporal and spectral speech signals. Hence, the model can effectively separate the speech signals and noise from an input signal. Two SNR-aware algorithms were proposed using CNN with the intention of improving the generalization capability and accuracy of these models. The first algorithm incorporates a Multi-Task Learning (MTL) framework. The noisy speech signal is fed as input to the model. Given the input, the algorithm primarily restores noise-free speech signals. Then, the SNR level is estimated for the processed clean speech signals. The second algorithm is based on SNR adaptive de-noising. The algorithm initially computes the SNR level. Then, based on the calculated SNR level, a SNR-dependent CNN model is chosen for reducing the noise.. It was found that max-pooling is not required here for speech enhancement due to its reduced capability in representing complex speech patterns. It is justified from the results that the two proposed SNR-aware CNN models outperform the deep neural networks in terms of standardized objective evaluations, provided the number of layers and nodes are defined to be the same. Additionally, the SNR-aware CNN models possess enhanced denoising potential even with unseen SNR levels. This potrays promising robust potential for real-world applications. Most recently, in 2017, another CNN model was proposed towards complex spectrogram enhancement in order to solve the difficulty in phase estimation [71]. The proposed model identifies clean real and imaginary (RI) spectrograms from noisy spectrograms. These restored RI spectrograms are then utilized to generate enhanced speech waveforms. These waveforms possess phase information with high accuracy. Objective function was formulated using Multi-Metric Learning (MML) criterion such that more than one metric is deemed. The main idea behind MML is that any signal representation can be portrayed as a function of RI spectrograms. With optimal selection of β, MML can boost multiple objective metrics (log-spectral distortion(LSD) and segmental signal-to-noise ratio (SSNR)) concurrently. The lift in the performance can be justified by considering MML as a pseudo layer over the original objective function. This process is believed to improve the generalization capability of the original model.

6.4 Using Bayesian with ICA for speech enhancement

In 2015, a Bayesian single-channel speech enhancement algorithm was proposed for the independent component analysis (ICA) domain to exploit sparseness in speech [72]. Generally, de-noising in the ICA domain is based on the unrealistic assumption that the noise components in this domain are uncorrelated. In the proposed approach, such assumption is not made and the general scenario is considered. The methodology is composed of two components. They include:

(1) estimator of maximum a posteriori (MAP) for speech coefficients in ICA domain, which is further adopted to estimate enhanced speech in the time domain
(2) transformation of data to ICA domain, learned from speech training data and then used in previous step

The proposed method was trained and experimented with speech keywords like commands for car navigation. This method demonstrated a substantial improvement in the de-noising performance with respect to SNR and distortion in enhanced signals when compared to the real time noisy speech signals from car, street, office and industrial environments.
Table 6.4 presents the summary of the various kinds of neural networks in the field of Speech enhancement.

6.5 Optimization techniques for speech enhancement

This section reviews a few prominent and recent optimization techniques with regard to speech enhancement. All the optimization techniques mentioned here consider that such a dual channel enhancement is used where one channel is for pure noise while the other is dedicated to speech distorted by noise.

6.5.1 Learning-based PSO (LPSO). In 2010, Learning-based Particle Swarm Optimization (LPSO), which is an improved stochastic optimization algorithm, was introduced to devise an adaptive filter for dual-channel speech enhancement application [73]. The search of region around the best solution is performed through dynamic search method. The algorithm then involves adaptive local search on each particle. During the process, sub-swarms exchange the best solutions at regular intervals

Table 5. Pros and cons of Adaptive noise cancelerPros and cons of machine learning methods

	Pros	Cons
NN[67]	• Low computational complexity • Less processing delay	• Needs improvement in accuracy • Not exceptionally good in terms of generalization performance to unpredictable conditions
DNN	• Better performance than SNN-based and L-MMSE methods [68] • Remarkable improvements in both objective and subjective metrics when compared with conventional MMSE based technique • Quite effective in handling real-world distorted noisy speech in various languages and across varying recording conditions not observed during DNN training[69] • Effective suppression of highly non-stationary noise, which is usually difficult to deal.[69]	• Improvement needed in generalization capability of DNN towards unseen noise [69] • Demand for large training set to provide good coverage of different acoustic environments such as speaker and language variations[69]
CNN	• Higher performance than DNN[70] • Efficient in handling local spectral and temporal structures of speech signals.[70] • Effective decomposition of the speech and noise signals from the noisy input signals[70] • Lack of necessity for Max pooling[70]. • Enhanced de-noising performance with unseen SNR levels[70] • Promising approach for real world applications[71][70] • Accurate phase information in enhanced signal[71] • Enhancement in multiple metrics simultaneously[71].	• computationally expensive approach[71]

through subpopulation strategy. The simulation results prove that the proposed LPSO algorithm outperforms the Standard Particle Swarm Optimization (SPSO), Genetic Algorithms (GA) and gradient-based NLMS algorithm with respect to SNR and stability. During another attempt in 2010, a hybrid optimization algorithm was suggested to boost the distorted speech signals in the framework of dual-channel speech enhancement[74]. The proposed hybrid algorithm θ-SSPSO combines the conventional θ-PSO and the Shuffled Sub-Swarms Particle Optimization (SSPSO) technique to exploit the advantages of both the algorithms. Experimental results reveal that the θ-SSPSO algorithm is highly effective in terms of global convergence for adaptive filters. Global convergence helped in achieving improved noise suppression in the candidate speech signal. θ-PSO algorithm, though characterizes a better optimization performance than the SPSO in the case of simple problems, but gets trapped in local optima when dealing with complex multi-objective functions. SSPSO overcomes this issue by increasing the diversity of particles in the search space thereby avoiding the local optima.

6.5.2 BAT Algorithm. A population-based meta-heuristic approach called the Bat Algorithm (BA), motivated by the hunting behavior of bats, was devised [75]. BA is rooted on the echolocation behavior of microbats. The algorithm adopts frequency tuning to elevate the diversity of the solutions in the population. It also implements the automatic zooming characteristic of bats such as the pulse emission rate and loudness on approaching the prey as the automatic adjustment capability in the algorithm. The capability attempts to balance exploration and exploitation during the search process by adapting from exploration to exploitation with the approaching of global optimality. This algorithm, being the first attempt to balance these important components, justifies itself to be a very efficient optimization technique when compared to other meta-heuristic algorithms [76]. Yet another attempt using Bat Algorithm (BA) towards dual channel speech enhancement systems was put forth in [77]. In this approach, BA is utilized in determination of the weights for the adaptive filter. The methodology initially involves segmenting the input signals into frames. Then, the objective function is formulated as the mean square error between the distorted speech and the estimated noise signal in each frame. Then, the optimization of the filter co-efficient is done through BA. Results justify that BA portrays an improved performance when compared SPSO algorithm in terms of improved quality and intelligibility in the enhanced speech.
In 2016, simulation results based on BA were compared with those of standard, accelerated PSO, gravitational search algorithm (GSA) and hybrid PSOGSA- based speech enhancement algorithms [78]. Results evidently demonstrate the potential of the meta-heuristic BA over the other algorithms pertaining to enhancement of speech signals.

6.5.3 Modified BAT algorithm. In 2015, an enhancement was formulated to the original BA [79]. The improvement pertains to adopting fuzzy system to dynamically adapt its parameter such as wavelength, loudness, low frequency and high frequency unlike the usual parameter tuning, which is performed based on trial and error. The proposed modification to BA is shown in the figure below. The results provide a comparison of the proposed modified algorithm with the original BA and Genetic Algorithms, depicting the effectiveness of the modification. Tests were also carried out with benchmark mathematical functions to demonstrate the potential of the proposed enhancement.

6.5.4 GSA Algorithm. In 2014, an optimization algorithm rooted on the law of gravity known as GSA was put forth [80]. It is a population-based algorithm. Agents (individuals) are regarded

as objects and their performance is estimated through masses. Objects attract each other due to force of gravity. Objects with heavier mass have high gravitational force and tend to attract objects with lower mass. Hence objects interact with each other by means of gravitational force. The objects with heavier mass are candidates for good solutions. These objects tend to move slower than the lighter ones, thereby improving exploitation. GSA achieves improved PESQ scores when compared to SPSO algorithm. Although SPSO finds good solutions, it suffers from the problem of local optimum. GSA yields better quality and intelligibility in the enhanced speech signals than that provided by SPSO.

6.5.5 PSOGSA algorithm. In 2015, a hybrid PSOGSA was presented to enhance the noise distorted speech signals in dual channel systems [81]. Each agent in the swarm, representing the filter coefficients is deemed as a candidate solution. PSOGSA is adopted to optimize these coefficients of adaptive filter. The performance of PSOGSA excelled the performance of that of GSA and SPSO. The hybrid algorithm possesses the advantage of exploration and exploitation capabilities of GSA and PSO respectively. Therefore, PSOGSA suppresses the background unwanted noise signals of the noisy input speech signals more effectively.
Table*6.5.5* shows the highlights of optimization methods reviewed. Having presented on the neural network based approaches towards speech enhancement, the following sections deal with the multi-modal approaches for improvement of speech signals.

7. MULTIMODAL APPROACHES TO SPEECH ENHANCEMENT

In recent years, it was investigated to check if information on visual speech could be utilized to boost the audio speech signals, contaminated by noise.
An attempt in 2002 was made to investigate the impact of enhancing noisy audio features using audio-visual speech data on Automatic Speech Recognition (ASR) systems [82]. The speech signals are enhanced through application of linear filters on concatenated audio and visual features. Improvements were noticeable to a large extent while using these features instead of the original noisy audio features be it for small or large vocabulary recognition. However, on comparison with the audio-visual discriminant feature fusion, the proposed approach yielded lower performance. The inferior performance can be attributed to the simplicity of linear filter utilized for enhancing, the non-stationary property of the noise considered and the characteristics of the proposed enhancement approach itself. The proposed method demands the resulting audio features to approximate the clean audio and hence the speech information in the visual, that provides complementary information, is not fully exploited. On the other hand, audio-visual feature fusion exploits the complementary information better by seeking the speech discriminant projection of joint audio-visual data, with no limitation to restrict within the original audio space. The adoption of non-linear systems, instead of linear ones for audio-visual speech enhancement can improve the recognition performance better.
In 2009, visually-derived Wiener filter was put forth for speech Enhancement by Almajai and Milner [83]. It uses both audio-visual correlation and the robustness of visual features to noise, for providing estimates of the clean speech. Investigation on RMS error reveals the estimation of both of these properties to be relatively robust to noise. Although it was effective in suppression of noise, but the visually-derived Wiener filter also introduced distortion onto the speech signal during the filtering process.
Almajai in [42] made use of visual speech information within a Wiener filter to improve the noisy speech signal. The approach reports to achieve improvement in the noisy speech, especially by reducing noise intrusiveness at the cost of signal distortion. The basic idea behind is the existence of correlation between the audio and video speech signals, facilitating the estimation of filterbank features from the visual features. The initial investigation of this approach reported various findings. Primarily, the correlation metric is higher when estimated within phonemes than globally across all speeches. Then investigation on measurements of filterbank estimation errors, subjective and objective tests reveals that the proposed method is relatively insensitive to phoneme decoding errors. In other words, only a very little difference was observed in the filterbank estimation errors when decoding accuracy decreased from 100% to 30%. Results also bring to notice that the estimation of spectral features from visual features is limited by the audio-visual correlation and also by the amount of speech information conveyed in lip movement. To illustrate an example, spectral details such as harmonic structure cannot be determined from the visual features. Hence, this places a limit on the level of spectral detail that can be extracted from the visual features. However, analysis has shown that coarse filterbank estimates are sufficient to enable speech enhancement to an extent.
In 2013, a two-stage multimodal speech enhancement framework, utilizing audio and visual information was proposed [84]. The input noise contaminated speech signals, obtained from microphone array is initially pre-processed through visually derived Gaussian Mixture Regression based Weiner filter, using visual speech information elicited by means of Semi Adaptive Appearance Models (SAAM) based lip tracking approach. Subsequently, the pre-processed speech signals are improved further though Transfer Function Generalized Sidelobe Canceller (TFGSC) approaches. The two-stage system is a promising solution in challenging noisy scenarios. Results provide a favorable outlook on the framework to be used in difficult noisy environments. The system is then extended to incorporate fuzzy logic to demonstrate proof of concept for an envisaged autonomous, adaptive, and context aware multimodal system[85]. The drawbacks in the system are that Weiner filters used are very basic and it uses less complex GMM for speech estimation.
In 2017, an audio-visual deep CNN (AVDCNN) Speech Enhancement model [86], that incorporates audio and visual streams into a unified network model, was put forth. The proposed approach was motivated by multi-modal Learning, incorporating data from different modalities and the proven potential of CNN in Speech Enhancement related tasks. Individual CNNs are primarily adopted to process the audio and visual data. After that, they are fused into unified network to produce enhanced speech at the output. Training of the proposed model is done end-to-end and back propagation learning is used for tuning the parameters. Results assessed based on five objective functions demonstrate that the AVDCNN excels the audio only CNN and other traditional SE methods, justifying the effectiveness of incorporating visual information into the process of speech signal enhancement. The coming table summarize the advantages and the disadvantages of the previous multimodal section.

8. CONCLUSION AND FUTURE WORK

In this paper, a survey of how researchers have tackled the issue of speech enhancement over the years have been presented. The

Table 6. Highlights of optimization methods for enhancing speech

	Highlights
LPSO[73]	• Higher performance when compared to SPSO, GA, and gradient-based NLMS algorithm in terms of SNR improvement and stability.
θ-PSO[74]	• Combination of pros of both algorithms, θ-PSO and SSPSO • Quite effective in achieving global convergence for adaptive filters • Better suppression of noise in the input speech signal • Increased diversity of particles in the search space to avoid getting caught in local optima. • Better than standard PSO, θ-PSO, and SSPSO with respect to convergence rate and SNR improvement • Possibility of getting trapped in local minima while dealing with complex or multi-mode functions.
GSA[80]	• Improved PESQ scores when compared to SPSO algorithm
PSOGSA[81]	• Better than GSA and SPSO
BAT[78]	• Better improved quality and intelligibility of enhanced speech than PSO, SPSO, APSO, GSA, PSOGSA
Modified BAT[79]	• Better than BAT and GA

Table 7. Advantages and dis-advantages of multimodal speech enhancement methods

	Pros	cons
Goecke et. al[82]	• Effective in enhancing the speech signal to improve speech recognition results	• Proposed method was not better than audio-visual discriminant feature fusion • speech information from the visual data is not fully exploited
Almajai & Milner[83]	• robust to noise effective in suppression of noise	• introduced distortion onto the speech signal during the filtering process.
Almajai & Milner[42]	• improvement in the noisy speech, especially by reducing noise intrusiveness • estimation of spectral features from visual features is limited	• signal distortion • coarse filterbank estimates are sufficient to enable speech enhancement to an extent but it not by a very effective margin
Abel and Hussain[84]	• promising solution in challenging noisy scenarios	• Weiner filters used are very basic • It uses less complex GMM for speech estimation.
Jen et.al[86]	• CNN approach is used • This AVDCNN excels the audio only CNN and other traditional SE methods	

earliest works done in this domain consist of the various kinds of spectral enhancement methods, statistical based algorithms and subspace enhancement methods. These have performed well under test conditions but in practical scenarios each comes with its own sets of drawbacks.

Adaptive noise cancellation is another popular domain in this regard. It has made itself an evergreen topic for research by being customizable through the use of machine learning techniques of optimization to tune its coefficients. Machine learning algorithms are quite vast in nature. It is not possible to cover them all within the scope of this paper. We have discussed a few prominent ones and enlisted the strong points of each.

Advances in the field of Artificial intelligence have yielded fruitful results in speech enhancement. Neural networks have proven to be a strong tool in this regard. After simple NN, came DNN which was stronger in results but showed poor real world generalization upon encountering noise and speech signals that were unseen to it during training phase. Then came the era of CNN, which has finally proven to be a reliable tool for generalization of real world noise cancellation problems. It can effectively deal with noise signals of all kinds, whether seen or unseen to it during training phase.

In future we will investigate and experimenting with optimization machine learning based filters for speech enhancement in real world scenarios.

Speech enhancement is the basis for all audio and communication devices. It is a technology that is rapidly growing by the day- and so it must for a technologically sound tomorrow.

9. REFERENCES

[1] Shishir Banchhor, Jimish Dodia, and Darshana Gowda. Gui based performance analysis of speech enhancement techniques. *International Journal of Scientific and Research Publications*, 3(9):1, 2013.

[2] Kumar K Ravi and PV Subbaiah. A survey on speech enhancement methodologies. *International Journal of Intelligent Systems and Applications*, 8(12):37, 2016.

[3] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[4] Hardik Panchmatia, Karan Gaikar, and Dharmesh Patel. Comparison of different speech enhancement techniques. *Imperial Journal of Interdisciplinary Research*, 2(5), 2016.

[5] Soumasunderaswari D and Prashanthini K. A survey on various multichannel speech enhancement algorithms. pages 254–255, 01 2015.

[6] Sunita Dixit and Dr MD Yusuf Mulge. Review on speech enhancement techniques. *International Journal of Computer Science and Mobile Computing, IJCSMC*, 3(8):285–290, 2014.

[7] A. Chaudhari and S. B. Dhonde. A review on speech enhancement techniques. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–3, Jan 2015.

[8] Devyani S Kulkarni, Ratnadeep R Deshmukh, and Pukhraj P Shrishrimal. A review of speech signal enhancement techniques. *International Journal of Computer Applications*, 139(14), 2016.

[9] V Sunnydayal, N Sivaprasad, and T Kishore Kumar. A survey on statistical based single channel speech enhancement techniques. *International Journal of Intelligent Systems and Applications*, 6(12):69, 2014.

[10] Jae S Lim and Alan V Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[11] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series : with engineering applications*. M.I. T. paperback series. Cambridge, Mass. Technology Press of the Massachusetts Institute of Technology, 1949.

[12] Aarti Singh. Adaptive noise cancellation. *Central Elektronica Engineering Research Institute, University of Dehli*, 2001.

[13] Nasser Mohammadiha, Timo Gerkmann, and Arne Leijon. A new linear mmse filter for single channel speech enhancement based on nonnegative matrix factorization. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 45–48. IEEE, 2011.

[14] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, volume 4, pages 208–211, Apr 1979.

[15] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), Apr 1979.

[16] Stefan J Mauger, Chris D Warren, Michelle R Knight, Michael Goorevich, and Esti Nel. Clinical evaluation of the nucleus® 6 cochlear implant system: Performance improvements with smartsound iq. *International journal of audiology*, 53(8):564–576, 2014.

[17] Tobias Goehring, Federico Bolner, Jessica JM Monaghan, Bas van Dijk, Andrzej Zarowski, and Stefan Bleeck. Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing research*, 344:183–194, 2017.

[18] Gianluca Monaci. On the modelling of multi-modal data using redundant dictionaries. 2007.

[19] Jon Driver. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381(6577):66, 1996.

[20] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[21] Mark T Wallace, GE Roberson, W David Hairston, Barry E Stein, J William Vaughan, and Jim A Schirillo. Unifying multisensory signals across time and space. *Experimental Brain Research*, 158(2):252–258, 2004.

[22] Shams Watkins, Ladan Shams, Sachiyo Tanaka, J-D Haynes, and Geraint Rees. Sound alters activity in human v1 in association with illusory visual perception. *Neuroimage*, 31(3):1247–1256, 2006.

[23] Artem Violentyev, Shinsuke Shimojo, and Ladan Shams. Touch-induced visual illusion. *Neuroreport*, 16(10):1107–1110, 2005.

[24] Jean-Pierre Bresciani, Franziska Dammeier, and Marc O Ernst. Vision and touch are automatically integrated for the perception of sequences of events. *Journal of vision*, 6(5):2–2, 2006.

[25] In Jean-Philippe Thiran, , Ferran Marqus, , and Herv Bourlard, editors, *Multimodal Signal Processing*, pages iv –. Academic Press, Oxford, 2010.

[26] S Lakshmikanth, KR Nataraj, and KR Rekha. Noise cancellation in speechsignal processing: A review. *International Journal of Advanced Research in Computer and Communication Engineering*, (1), 2014.

[27] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[28] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza. *Nonlinear Speech Enhancement: An Overview*, pages 217–248. Springer Berlin Heidelberg, 2007.

[29] Sunil Kamath and Philipos C. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *ICASSP*, page 4164. IEEE, 2002.

[30] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11(2):215 – 228, 1992.

[31] S Ogata and Tetsuya Shimamura. Reinforced spectral subtraction method to enhance speech signal. In *TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, volume 1, pages 242–245. IEEE, 2001.

[32] Nathalie Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on speech and audio processing*, 7(2):126–137, 1999.

[33] Hwai-Tsu Hu, Fang-Jang Kuo, and Hsin-Jen Wang. Supplementary schemes to spectral subtraction for speech enhancement. *Speech Communication*, 36(3?4):205 – 218, 2002.

[34] Navneet Upadhyay and Abhijit Karmakar. An improved multi-band spectral subtraction algorithm for enhancing speech in various noise environments. *Procedia Engineering*, 64:312–321, 2013.

[35] C. Yu and L. Su. Speech enhancement based on the generalized sidelobe cancellation and spectral subtraction for a microphone array. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 1318–1322, Oct 2015.

[36] L. Cao, T. q. Zhang, H. x. Gao, and C. Yi. Multi-band spectral subtraction method combined with auditory masking properties for speech enhancement. In *2012 5th International*

Congress on Image and Signal Processing, pages 72–76, Oct 2012.

[37] Yu Cai and Chaohuan Hou. Subband spectral-subtraction speech enhancement based on the dft modulated filter banks. In *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*, volume 1, pages 571–574. IEEE, 2012.

[38] Prabhakaran G., Indra J., and Kasthuri N. Tamil speech enhancement using non-linear spectral subtraction. In *2014 International Conference on Communication and Signal Processing*, pages 1482–1485, April 2014.

[39] Md T Islam, C Shahnaz, and SA Fattah. Speech enhancement based on a modified spectral subtraction method. In *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1085–1088. IEEE, 2014.

[40] Shambhu Shankar Bharti, Manish Gupta, and Suneeta Agarwal. A new spectral subtraction method for speech enhancement using adaptive noise estimation. In *Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on*, pages 128–132. IEEE, 2016.

[41] Guo-Hong Ding, Taiyi Huang, and Bo Xu. Suppression of additive noise using a power spectral density mmse estimator. *IEEE Signal Processing Letters*, 11(6):585–588, June 2004.

[42] I. Almajai and B. Milner. Visually derived wiener filters for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1642–1651, Aug 2011.

[43] Marco Jeub and Peter Vary. Binaural dereverberation based on a dual-channel wiener filter with optimized noise field coherence. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4710–4713. IEEE, 2010.

[44] MA Abd El-Fattah, Moawad Ibrahim Dessouky, Salah M Diab, and Fathi El-Sayed Abd El-Samie. Speech enhancement using an adaptive wiener filtering approach. *Progress In Electromagnetics Research M*, 4:167–184, 2008.

[45] Marwa A Abd El-Fattah, Moawad I Dessouky, Alaa M Abbas, Salaheldin M Diab, El-Sayed M El-Rabaie, Waleed Al-Nuaimy, Saleh A Alshebeili, and Fathi E Abd El-Samie. Speech enhancement with an adaptive wiener filter. *International Journal of Speech Technology*, 17(1):53–64, 2014.

[46] Amart Sulong, Teddy Surya Gunawan, Othman O Khalifa, Mira Kartiwi, and Eliathamby Ambikairajah. Speech enhancement based on wiener filter and compressive sensing. *Indonesian Journal of Electrical Engineering and Computer Science*, 2(2):367–379, 2016.

[47] Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, 1985.

[48] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, Dec 1984.

[49] Olivier Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE transactions on Speech and Audio Processing*, 2(2):345–349, 1994.

[50] Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement *IEEE Transactions on speech and audio processing*, 3(4),251–266, 1995.

[51] Yi Hu and Philipos C. Loizou. A subspace approach for enhancing speech corrupted by colored noise. In *ICASSP*, pages 573–576. IEEE, 2002.

[52] S. Surendran and T. K. Kumar. Perceptual subspace speech enhancement with ssdr normalization. In *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, pages 1–6, Jan 2016.

[53] F. Jabloun and B. Champagne. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 11(6):700–708, Nov 2003.

[54] Wang Guang Yan, Geng Yan Xiang, and Zhao Xiao Qun. A signal subspace speech enhancement method for various noises. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(2):726–735, 2013.

[55] Chengli SUN, Jianxiao XIE, and Yan LENG. A signal subspace speech enhancement approach based on joint low-rank and sparse matrix decomposition. *Archives of Acoustics*, 41(2), 2016.

[56] Sultana Hifrin, Nayan Bharali Palash, and Sharmah Uzzal. General report on speech recognition using pattern classification methods. *Communication, Cloud and Big Data: Proceedings of CCB 2014*, 2014.

[57] S Lakshmikanth, KR Nataraj, and KR Rekha. Noise cancellation in speechsignal processing: A review. *International Journal of Advanced Research in Computer and Communication Engineering*, (1), 2014.

[58] Bernard Widrow and Samuel D Stearns. *Adaptive signal processing*, volume 15. Prentice-hall Englewood Cliffs, NJ, 1985.

[59] Prajna Kunche and KVVS Reddy. *Metaheuristic Applications to Speech Enhancement*. Springer, 2016.

[60] Simon Haykin. *Adaptive Filter Theory (3rd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[61] E Hari Krishna, M Raghuram, K Venu Madhav, and K Ashoka Reddy. Acoustic echo cancellation using a computationally efficient transform domain lms adaptive filter. In *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, pages 409–412. IEEE, 2010.

[62] Huang Guopin, Zhao Wei, and Zhang Qin. Improvement of audio noise reduction system based on rls algorithm. In *Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on*, pages 964–968. IEEE, 2013.

[63] Rakesh, Pogula and Kumar, T Kishore, A Novel RLS Based Adaptive Filtering Method for Speech Enhancement *International Journal of Electrical, Computer, Electronics and Communication Engineering, World Academy of Science, Engineering and Technology* 9(2), 624–628, 2015

[64] Mohamed Djendi, Rahima Henni, and Akila Sayoud. A new dual forward bss based rls (dfrls) algorithm for speech enhancement. In *Engineering & MIS (ICEMIS), International Conference on*, pages 1–4. IEEE, 2016.

[65] M. M. Dewasthale, R. D. Kharadkar, and M. Bari. Comparative performance analysis and hardware implementation of adaptive filter algorithms for acoustic noise cancellation. In *2015 International Conference on Information Processing (ICIP)*, pages 124–129, 2015.

[66] Jyoti Dhiman, Shadab Ahmad, and Kuldeep Gulia. Comparison between adaptive filter algorithms (lms,

nlms and rls). *International Journal of Science, Engineering and Technology Research (IJSETR)*, 2(5):1100–1103, 2013.

[67] Tobias Goehring, Federico Bolner, Jessica JM Monaghan, Bas van Dijk, Andrzej Zarowski, and Stefan Bleeck. Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing research*, 344:183–194, 2017.

[68] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2014.

[69] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(1):7–19, 2015.

[70] Szu-Wei Fu, Yu Tsao, and Xugang Lu. Snr-aware convolutional neural network modeling for speech enhancement. In *INTERSPEECH*, pages 3768–3772, 2016.

[71] Szu-Wei Fu, Ting-yao Hu, Yu Tsao, and Xugang Lu. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning, 2017.

[72] Hong, Liang and Rosca, Justinian and Balan, Radu. Bayesian single channel speech enhancement exploiting sparseness in the ICA domain. *Signal Processing Conference, 2004 12th European*, 1713–1716.IEEE, 2004

[73] L Badri Asl and Masoud Geravanchizadeh. Dual-channel speech enhancement based on stochastic optimization strategies. In *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, pages 229–232. IEEE, 2010.

[74] Sina Ghalami Osgouei and Masoud Geravanchizadeh. Dual-channel speech enhancement based on a hybrid particle swarm optimization algorithm. In *Telecommunications (IST), 2010 5th International Symposium on*, pages 873–877. IEEE, 2010.

[75] Xin-She Yang. *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[76] K Prajna, G Sasibhushana Rao, KVVS Reddy, and R Uma Maheswari. Application of bat algorithm in dual channel speech enhancement. In *Communications and Signal Processing (ICCSP), 2014 International Conference on*, pages 1457–1461. IEEE, 2014.

[77] Prajna Kunche and KVVS Reddy. Speech enhancement based on bat algorithm (ba). In *Metaheuristic Applications to Speech Enhancement*, pages 91–110. Springer, 2016.

[78] Prajna Kunche and KVVS Reddy. *Metaheuristic Applications to Speech Enhancement*. Springer, 2016.

[79] Jonathan Pérez, Fevrier Valdez, and Oscar Castillo. Modification of the bat algorithm using fuzzy logic for dynamical parameter adaptation. In *Evolutionary Computation (CEC), 2015 IEEE Congress on*, pages 464–471. IEEE, 2015.

[80] K Prajna, GSB Rao, KVVS Reddy, and R Uma Maheswari. A new approach to dual channel speech enhancement based on gravitational search algorithm (gsa). *International Journal of Speech Technology*, 17(4):341–351, 2014.

[81] Prajna Kunche, G Sasi Bhushan Rao, KVVS Reddy, and R Uma Maheswari. A new approach to dual channel speech enhancement based on hybrid psogsa. *International Journal of Speech Technology*, 18(1):45–56, 2015.

[82] R. Goecke, G. Potamianos, and C. Neti. Noisy audio feature enhancement using audio-visual speech data. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2025–II–2028, May 2002.

[83] Ibrahim Almajai and Ben Milner. Enhancing audio speech using visual speech features. In *In:proc.Interspeech, Brighton, UK*, 2009.

[84] Andrew Abel and Amir Hussain. *Cognitively Inspired Audiovisual Speech Filtering: Towards an Intelligent, Fuzzy Based, Multimodal, Two-Stage Speech Enhancement System*, chapter A Two Stage Multimodal Speech Enhancement System, pages 35–51. Springer International Publishing, Cham, 2015.

[85] Andrew Abel and Amir Hussain. *Towards Fuzzy Logic Based Multimodal Speech Filtering*, pages 75–90. Springer International Publishing, Cham, 2015.

[86] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Jen-Chun Lin, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement based on multimodal deep convolutional neural network. *arXiv preprint arXiv:1703.10893*, 2017.