Nonlinear Speech Enhancement: An Overview

A. Hussain[1], M. Chetouani[2], S. Squartini[3], A. Bastari[3], and F. Piazza[3]

[1] Department of Computing Science & Mathematics, University of Stirling, Stirling FK9 4LA, Scotland, UK
[2] Laboratoire des Instruments et Systemes d'Ile-De-France Université Paris Pierre and Marie Curie, 4 Place Jussieu, 75252 Paris Cedex 05, France
[3] Dipartimento di Elettronica, Intelligenza Artificiale e Telecomunicazioni Università Politecnica delle Marche Via Brecce Bianche 12, I-60121 Ancona, Italy
ahu@cs.stir.ac.uk, mohamed.chetouani@upmc.fr,
sts@deit.univpm.it, a.bastari@univpm.it, upf@deit.univpm.it

Abstract. This paper deals with the problem of enhancing the quality of speech signals, which has received growing attention in the last few decades. Many different approaches have been proposed in the literature under various configurations and operating hypotheses. The aim of this paper is to give an overview of the main classes of noise reduction algorithms proposed to-date, focusing on the case of additive independent noise. In this context, we first distinguish between single and multi channel solutions, with the former generally shown to be based on statistical estimation of the involved signals whereas the latter usually employ adaptive procedures (as in the classical adaptive noise cancellation scheme). Within these two general classes, we distinguish between certain subfamilies of algorithms. Subsequently, the impact of nonlinearity on the speech enhancement problem is highlighted: the lack of perfect linearity in related processes and the non-Gaussian nature of the involved signals are shown to have motivated several researchers to propose a range of efficient nonlinear techniques for speech enhancement. Finally, the paper summarizes (in tabular form) for comparative purposes, the general features, list of operating assumptions, the relative advantages and drawbacks, and the various types of nonlinear techniques for each class of speech enhancement strategy.

Keywords: Single-channel/Multi-channel Speech Enhancement, Noise Reduction, Noise Cancellation, Microphone array, Non-linear techniques.

1 Introduction

The goal of speech enhancement systems is either to improve the perceived quality of the speech, or to increase its intelligibility [1-3]. There is a large variety of real world applications for speech enhancement in audio signal processing – for example, we experience the presence of degraded speech both in military and commercial communications, induced by different transmission channels (telephony) or produced in various noisy environments (vehicles, home-office etc.). Due to the growing interest in this subject, numerous efforts have been made over the past 20 years or so by the scientific community in order to find an effective solution for the speech enhancement problem. The different nature of interfering sounds, the admissible assumptions on the generating process of speech degradation and on the available observables, and the

Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.): WNSP 2005, LNCS 4391, pp. 217 – 248, 2007.

various operating conditions involved require us to make a preliminary distinction between the various speech enhancement approaches proposed to-date.

The principal aim of this work is to give an overview and a preliminary comparison of the main up to date techniques found in the literature for the noise reduction problem, focusing on non-linear techniques in particular. However, wherever required, some linear techniques are also outlined in order to better introduce certain non-linear extensions of interest. Because of the huge amount and diverse range of works reported in this field, we only introduce certain families of algorithms, considering the case when the noise is additive and independent of the clean speech. Note that speech de-reverberation and separation case studies, which can also be classified as speech enhancement problems, are not addressed here. The mathematics behind the reviewed methods is also omitted for lack of space, only the main formulae being introduced. Experimental numerical results and direct comparisons of the different introduced techniques are avoided firstly, because standard benchmark data have not been proposed or used in the literature, and secondly because the various reported methods all make use of different and individually optimized operating assumptions. Instead, the main details of the important reviewed methods will be summarized and compared in tabular form at the end of this paper.

Initially, we make a rough distinction between the different techniques by considering the number of available noisy speech channels: in the next paragraphs of this Section, the single channel speech enhancement problem and the multi channel speech enhancement problem are introduced; Section II further discusses the main nonlinear approaches developed for the monaural speech enhancement problem including supervised and unsupervised neural network based techniques. Section III addresses the two-channel/binaural speech enhancement case. Section IV reviews the general multi-channel non-linear speech enhancement case and finally, Section V presents the concluding summary by highlighting the main features, list of operating assumptions, the relative advantages and drawbacks and the various types of non-linear techniques for each class of speech enhancement strategy reviewed in this paper.

1.1 Basic Concepts

The degradation of the speech signal can be modeled, in a quite general manner, as follows:

$$y[k]=h[k]*s[k]+n[k]=s^h[k]+n[k] \tag{1}$$

where s^h is the observed degraded speech, s is the original signal to be recovered, and n is the additive noise, *h[k]* is the impulse response of the room where the sensor is placed and $*$ represents the convolution operator. Obviously, one can think of other types of degradation models that require specific enhancing methods. For the most part of this work, the convolutive term is not considered ($h[k]=\delta[k]$), with only the additive term $n[k]$ considered present. In addition, we shall consider background noise as interference in our studies, and the cases of speech separation (cocktail party problem), impulse or transient noises [1] will not be dealt with here due to space restrictions.

A first rough distinction between speech enhancement techniques can be made by looking at the number and type of observables available for:

1. single channel speech enhancement: where only one degraded version of the original speech is available and modeled by (1);
2. multi-channel speech enhancement: where the noisy observations are obtained from two or more sensors.

In conventional approaches [4], speech and noise signals are considered as unknown random processes and the objective is to perform an adequate statistical estimation of one random process (the speech signal) from the sum of speech and the noise. Such a task is hard because we have neither a precise statistical model of the signals nor a reliable measure to evaluate the effectiveness of the enhancement process. Moreover the non-stationarity of speech (and possibly of noise as well) [5] requires tracking of its time-varying statistical properties by means of adaptive solutions. Next, we present an overview of single-channel and multi-channel speech enhancement problems, followed by their non-linear extensions.

1.2 Brief Overview of the Single-Channel Speech Enhancement Problem

Spectral Subtraction (SS) [6] is probably the earliest and most well-known technique for single channel speech enhancement: it is often still used due to its efficacy and simplicity. In its most basic form (involving subtraction/filtering of power spectral density/amplitudes), the noise power spectral density is estimated, but the method introduces musical noise and other distortions in the recovered signal [1],[3]. Some interesting solutions involving nonlinear techniques have been proposed in the literature to overcome such drawbacks [1], [7]. However, as widely agreed, the best algorithm from this perspective is the one proposed by Ephraim and Malah [8-10] that is closely related to the pioneering work of McAulay and Malpass [11]. This is based on the minimum mean square error (MMSE) estimation of the speech spectrum in the logarithmic domain; and it is a natural extension of the one in the linear domain [8]. Further improvements have been achieved through the employment of better performing MMSE estimators, as in Xie and Compernolle [12], or by making the spectral subtraction procedure dependent on the properties of the human auditory system [13].

Another interesting derivative of SS is the signal subspace approach [14], [15] based on an estimation of the clean speech, as also done in the case of the Bayesian approach for speech enhancement using Hidden Markov Models (HMM) [16], [17]. Since this approach in its basic form is essentially linear and thus out of the intent of this work, it is not described in the following. HMM have also been successfully implemented in nonlinear estimation frameworks [18], [19] where some speech data is assumed available for training. Other methods relying on the availability of a suitable training set have also been developed. Among these, we can cite the time domain and transform domain nonlinear filtering methods employing neural networks [20-26].

On the other hand, unsupervised single-channel speech enhancement techniques have received significant attention recently. Examples here include the Extended Kalman Filtering [15] [27-28] Monte-Carlo simulations [4], Particle filtering [4], [21], [30] and the Noise-Regularized Adaptive Filtering [15], [31] approaches, that can enable significant noise reduction even in difficult situations (involving noise non-gaussianity or system nonlinearity).

1.3 Brief Overview of the Multi-channel Speech Enhancement Problem

The multi-channel speech enhancement problem can be modelled as follows [1]:

$$y_m[k] = h_m[k] * s[k] + n_m[k] = s_m^h[k] + n_m[k] \tag{2}$$

where m is the sensor index. When $m = 2$ we refer to the so called binaural case if the spacing between the microphones is comparable to that between human's ears.

In the last years the scientific community has particularly focused its attention on multichannel techniques, as they virtually provide remarkable outcomes on the single channel ones. As highlighted in some recent works [3], using a single channel it is not possible to improve both intelligibility and quality of the recovered signal at the same time. Quality can be improved at the expense of sacrificing intelligibility. A way to overcome this limitation is to add some spatial information to the time/frequency information available in the single channel case. We can get this additional information using two or more channel of noisy speech.

Adaptive noise cancellation [32-33] can be viewed as a particular case of the multi-channel speech enhancement problem. Indeed, we have two observables, the noisy speech and the reference noise, and the goal is to get an enhanced output speech adaptively according to the scheme in Fig.1. Classical methods based on full-band multi-microphone noise cancellation implementations can produce excellent results in anechoic environments with localized sound radiators, however performance deteriorates in reverberant environments. Adaptive sub-band processing has been found to overcome these limitations [34]. The idea of involving sub-band diverse processing to take account of the coherence between noise signals from multiple sensors has been implemented as part of the so-called Multi-Microphone Sub-Band Adaptive (MMSBA) speech enhancement system [35-38].

The main limitation of these linear approaches is that they are not able to deal effectively with non-gaussianity of the involved signals or the non-linear distortions arising from the electro-acoustic transmission systems. As a result, several nonlinear approaches have been proposed to-date mainly employing Neural Networks (NN) and Volterra Filtering (VF), see for example,[39-42], [25]. Such non-linear processing approaches have also been successfully implemented within the MMSBA architecture, as will be highlighted later on.

If available, more than two microphones (resulting in a microphone array) can be used in order to achieve better performance for noise reduction. The most common approaches here are represented by the delay-and-sum array and the adaptive beamformer [43]. Among the large variety of linear approaches that have appeared in the literature so far for speech enhancement, some nonlinear microphone arrays have also been proposed [44-46], which seem to exhibit relevant performance improvements with respect to their linear counterparts. Another interesting nonlinear approach in the microphone array area is represented by the idea of estimating the log spectra of involved signals (as in the single channel case), taking advantage of the availability of more sensors [47-51].

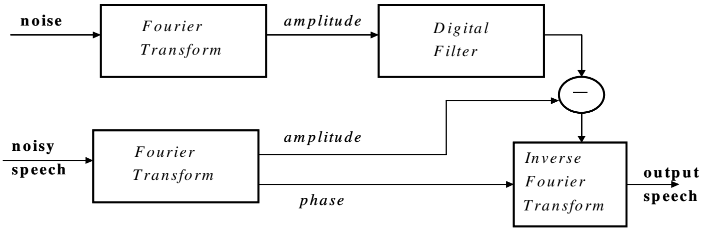

Fig. 1. Spectral Subtraction when two microphones are available (subtraction of amplitudes)

2 Nonlinear Monaural Speech Enhancement

In this section some well-known nonlinear methods for single-channel speech enhancement are outlined, without any intention of being exhaustive. Indeed, many other relevant contributions are present in the literature: our attempt here is to highlight the main approaches that have received much attention recently.

2.1 Spectral Subtraction (SS): Linear and Nonlinear Methods

Spectral Subtraction is a simple and effective method for reduction of stationary background noise [1-2], [6]. The processing is done on a frame-by-frame basis in the frequency domain. Speech and noise are assumed to be uncorrelated. The estimated speech short-time magnitude $\left|\hat{S}(\omega)\right|$ is obtained by subtracting from the noisy speech short-time magnitude |Y(ω)| a noise spectral magnitude estimate |Ñ(ω)| computed during speech pauses. This is what is essentially depicted in Fig.1, keeping in mind that the two-microphone SS scheme is equivalent to the monaural case with an effective Voice Activity Detector to allow the estimation of noise statistics during noise-alone periods. Taking into account the power spectral subtraction case study, we have

$$\left|\hat{S}(\omega)\right|^2 = \begin{cases} \left|Y(\omega)\right|^2 - \left|\hat{N}(\omega)\right|^2 & if \;\; \left|Y(\omega)\right|^2 > \left|\hat{N}(\omega)\right|^2 \\ 0 & otherwise \end{cases} \tag{3}$$

where $\left|\hat{N}(\omega)\right|^2$ is the noise power spectral estimate. The phase of noisy speech is left unchanged, so the enhanced signal in time domain is obtained as:

$$\hat{s}(k) = IFFT\left[\left|\hat{S}(\omega)\right| e^{j\arg(Y(\omega))}\right]. \tag{4}$$

Subtractive-type algorithms can be studied using a second approach termed *filtering of noisy speech*, involving the use of a time-varying linear filter dependent on the characteristics of the noisy signal spectrum and on the estimated noise spectrum. The noise suppression process becomes a product of the short-time spectral magnitude of the noisy speech $\left|Y(\omega)\right|$ with a gain function $G(\omega)$ as follows:

$$\left|\hat{S}(\omega)\right| = G(\omega)\left|Y(\omega)\right| \qquad with \quad 0 < G(\omega) < 1 . \tag{5}$$

In terms of power spectral densities and considering (3) we have:

$$G(\omega) = \sqrt{1 - \left(\left|\hat{N}(\omega)\right|^2 \Big/ \left|Y(\omega)\right|^2\right)} = \sqrt{R_{post}(\omega) / \left(1 + R_{post}(\omega)\right)} \tag{6}$$

that is constrained to be null if the estimated noise power level is superior to that of the noisy speech. $R_{post}(\omega) = \left(\left|Y(\omega)^2\right| / \left|N(\omega)^2\right|\right) - 1$ is the a posteriori SNR. In other words, such a subtractive scheme results in emphasizing the spectral components proportionally to the amount by which they exceed noise. As can be seen in (6), $G(\omega)$ can be written as a function of the a-posteriori SNR, and many different rules, namely suppression curves, have been proposed so far. Their aim is to make the application of $G(\omega)$ more flexible in order to reduce the effect of musical noise that is characteristic of the classical SS approach [1-2], [6]. From this perspective, an interesting solution has been proposed as the so-called nonlinear SS [7], according to which a nonlinear estimation of noise power spectral density $\left|\hat{N}(\omega)\right|_{nl}^2$ is used in (3) as follows:

$$\left|\hat{N}(\omega)\right|_{nl}^2 = \Phi\left(\max_{over\, M\, frames}\left(\left|\hat{N}(\omega)\right|^2\right), R_{post}(\omega), \left|\hat{N}(\omega)\right|^2\right) \tag{7}$$

where $\Phi(.)$ is the nonlinearity involved in the estimation process. A possible formulation for this is:

$$\Phi\left(\max_{over\, M\, frames}\left(\left|\hat{N}(\omega)\right|^2\right), R_{post}(\omega)\right) = \frac{\max_{over\, M\, frames}\left(\left|\hat{N}(\omega)\right|^2\right)}{1 + \gamma R_{post}(\omega)} \tag{8}$$

with γ being a design parameter. Equation (8) says that as the SNR decreases the output of the nonlinear estimator approaches the maximum value of noise spectrum over *M* frames, and as SNR increases it approaches zero. One can consider more complicated $\Phi(.)$, depending also on $\left|\hat{N}(\omega)\right|^2$, which can be useful if one is interested in over-subtraction for example.

2.2 The Ephraim-Malah SS Algorithm and Some of Its Variants

The Ephraim Malah algorithm [8-10] has received much attention by the scientific community. This is mainly due to its ability to achieve a highly satisfying overall quality of the enhanced speech which is appreciatively artifacts-free, and these characteristics makes it suitable for practical implementations in digital hearing aids. Such an approach has been down to outperform the conventional SS schemes as it is based on an estimation of the short-time spectral amplitude (STSA) of the speech signal. The same is also the case with the Soft-Decision Noise Suppression filter of McAulay and Malpass [11] where the STSA estimator is derived from an optimal (in the Maximum-Likelihood sense) variance estimator. In Ephraim and Malah (1985), an MMSE (minimum mean square error) STSA estimator is derived and applied in a SS scheme. The basic assumptions are the statistical independence of speech and noise, along with the spectral components of each of these two processes considered as zero mean statistically independent Gaussian random variables.

As pointed out in several papers, the main difference between the two STSA based approaches, i.e. [8] and [11], is that the former is able to yield colourless residual noise, whereas musical noise is still present after processing the observable through the latter procedure. In the following only the main formulae constituting the Ephraim-Malah noise suppressor are reported. Omitting the time and frequency indexes (l,ω) in order to shorten the notation, the suppression curve $G(l,\omega)$ to be applied to the short-time spectrum value $|Y(l,\omega)|$ can be expressed as:

$$G(l,\omega)=\frac{\sqrt{\pi}}{2}\cdot\sqrt{\left(\frac{1}{1+R_{post}}\right)\left(\frac{R_{prio}}{1+R_{prio}}\right)}\cdot M\left(\left(1+R_{post}\right)\left(\frac{R_{prio}}{1+R_{prio}}\right)\right) \tag{9}$$

where $M(.)$ is the nonlinearity based on 0th and 1st order Bessel functions:

$$M(\theta)=\exp\left(-\frac{\theta}{2}\right)\left[(1+\theta)I_0\left(\frac{\theta}{2}\right)+\theta I_1\left(\frac{\theta}{2}\right)\right]. \tag{10}$$

The formulations of the a-priori SNR and a-posteriori SNR respectively (for each value of the time and frequency indexes) are given below:

$$\begin{aligned} R_{post}(l,\omega)&=\left(\left|Y(l-1,\omega)\right|^2\Big/\left|\hat{N}(\omega)\right|^2\right)-1\\ R_{prio}(l,\omega)&=(1-\alpha)P\left[R_{post}(l,\omega)\right]+\alpha\frac{\left|G(l-1,\omega)Y(l-1,\omega)\right|^2}{\left|\hat{N}(\omega)\right|^2} \end{aligned} \tag{11}$$

with $P[x]=x$ if $x\geq 0$ and $P[x]=0$ otherwise. R_{prio} 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. Compared to other noise suppression rules based on averaging the short-time spectrum or on calculating the gain function over successive frames, one advantage of the Ephraim-Malah algorithm lies in the nonlinear averaging process. When the signal level is well above the noise level, the a-priori SNR becomes almost equivalent to the a-posteriori SNR with one frame delay, with the result that R_{prio} is no longer a smoothed SNR estimate (which is important for preventing the deterioration of the speech signal which is rather non-stationary).

The original version of the Ephraim-Malah rule does not take the signal presence uncertainty into account, in contrast to the procedure developed in [11]. This is a relevant aspect, since the speech signal is not always present in the noisy mixture and the energy of some voiced type spectral contributions is negligible in comparison to the corresponding noise.

An interesting generalization to the rule described by (9) has also been derived to address this problem. However it is not reported here, as it has been shown to behave similarly to the log-spectral estimator developed in [9]. Such an approach comprises a nonlinear spectral estimator performing the MMSE of the log-spectra. The underlying motivation is that a distortion measure based on the MSE of the log spectra is more subjectively meaningful than the counterpart based on the MSE of the common spectra. The spectral gain $G_{\log}(l,\omega)$ of the MMSE log spectral amplitude estimator is

$$G_{\log}(l,\omega)=\frac{R_{prio}}{1+R_{prio}}\cdot\exp\left[\frac{1}{2}\int_{\kappa(\omega)}^{+\infty}\frac{e^{-t}}{t}dt\right] \tag{12}$$

where R_{prio} and R_{post} are defined as above and the following holds:

$$\kappa(\omega)=\frac{R_{prio}}{1+R_{prio}}\left(1+R_{post}\right). \tag{13}$$

As observed by the authors, the rule (12) allows higher noise suppression, leaving unchanged the quality of the output speech with respect to the gain function in (9).

Further improvement in the performance achievable through this approach has been demonstrated in [12], who employ an empirical approach to yield a numerical solution to the MMSE estimate in the log spectral domain. Assuming that the speech and noise log spectra have normal distributions, it can be shown that the MMSE estimate of the speech log spectrum at certain time instant and frequency bin (l,ω) is a function of noisy observations and the probabilistic model parameter (mean and variance $\{\mu_s,\sigma_s,\mu_n,\sigma_n\}$). Such a function must be approximated, and the authors in [12] propose the novel use of a multi-layer perceptron (MLP) neural network. Monte Carlo simulations are used to get an adequate input/output training set for the network under the assumed statistics; and the approximation problem then turns out to be a curve fitting one by considering the MMSE estimation as a gain function. Considering the presence of a VAD to ensure the calculation of noise statistics during silence periods (even in slowly time-varying environments), assuming fixed and known the variance of the speech log spectra, and reformulating the parameter model after proper normalization, we can formulate the scheme of approximation of MMSE estimation as shown in Fig.2.

2.3 Overview of Supervised Neural Network Based Approaches

Other important nonlinear methods for single channel speech enhancement are proposed and analyzed in this section. These generally provide a suitable estimation of the clean speech signal, by means of nonlinear models in order to take into account the nonlinearities within the dynamic process determining the speech signal production. We shall consider here some techniques assuming the availability of a clean speech training data for the underlying nonlinear model. The classical techniques using Neural Networks as nonlinear filters mapping the noisy speech to clean speech in the time domain or in different domains [20], allow to get good estimations only

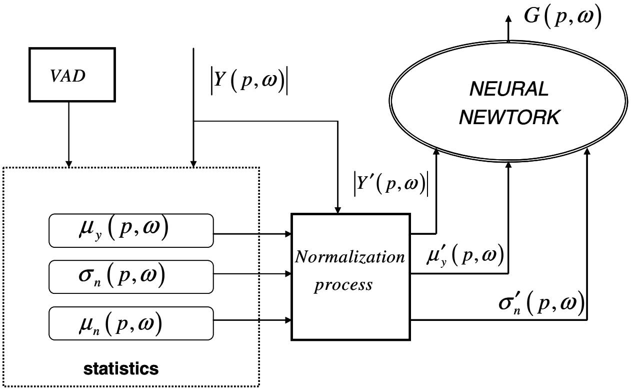

Fig. 2. Approximation process of the MMSE estimation in the log spectral domain

assuming speech and noise stationarity. A time variant model can be achieved by creating different fixed models for corresponding dynamical regimes of the signals and switching between these models during the speech enhancement process.

We start therefore from a straightforward neural extension of the work by Ephraim [16-17] which is represented by the principled switching method proposed by Lee [18], that incorporates the extended Kalman filtering approach (which will be discussed later). HMMs have been shown to be an effective tool in presence of signal uncertainty [16], due to their capability of dividing the received speech signal into various classes automatically. With reference to [18], each HMM state provides a maximum-likelihood estimate $\hat{s}(k)$ under the assumption that the windowed observation vector $\mathbf{y}(k)$ belongs to class *i*. The overall estimate is given by

$$\hat{s}(k)=\sum_i p\left(class_i\left|\mathbf{y}(k)\right.\right)\cdot\left[\hat{s}(k)\left|\mathbf{y}(k),class_i\right.\right] \tag{14}$$

where $p\left(class_i\left|\mathbf{y}(k)\right.\right)$ is the probability of being in class *i* given the window of noisy observations $\mathbf{y}(k)$ and the second term in the sum represents the maximum-likelihood estimate of the speech given class *i* and the data. The posterior class probability $p\left(class_i\left|\mathbf{y}(k)\right.\right)$ is easily calculated using standard forward-backward recursive formulas for HMMs. Alternatively, the estimate $\hat{s}(k)$ may be simply taken as the estimate for the single filter whose posterior class probability is maximum:

$$\hat{s}(k)=\left[\hat{s}(k)\left|class_m\right.\right] \quad with \quad p\left(class_m\left|\mathbf{y}(k)\right.\right)\geq p\left(class_i\left|\mathbf{y}(k)\right.\right) \quad \forall i\,. \tag{15}$$

The Extendend Kalman Filter (EKF) technique, involving an autoregressive model for each class, can be used to provide the maximum-likelihood estimation for speech. On purpose, a suitable set of clean speech data has to be employed to train the autoregressive neural models, whereas the speech innovations variance σ_n^2 can be estimated from the clean speech for each class.

A recent variant has been proposed [19] to the above approach of Lee et al [18]: wherein the nonlinear prediction model is based on a Recurrent Neural Network (RNN). The enhanced speech is the output of an architecture, namely RNPHMM (Recurrent Neural Predictive Hidden Markov Model), resulting from the combination of RNN and HMM. Similar to the previous approach [18], the unknown parameters are estimated by a learning algorithm derived from the Baum-Welch and RNN backpropagation algorithms.

As previously outlined Neural Networks can also be used as non-linear time domain filters, fed with the noisy speech signal to yield the estimate of the clean speech. The training is performed by using clean speech (from a known database) artificially corrupted to create noisy input data and presented to the network sliding the observation window over the available signal. The Tamura approach [22-23] is one of the oldest and most representative of this category: a four-layered neural network is used and trained for hetero-association, employing noisy speech signal patterns at the input and the corresponding noise free signal patterns at the output. Obtained results have been compared to those obtained with spectral subtraction through subjective listening tests, concluding that most listeners preferred the neural network filtered speech.

Another classical scheme is the one used in [24] where the noise signal is filtered through a feedforward network with a *M*-unit hidden layer and a single output unit, whose notation is used on e following. For each time instant k, the hidden unit computes the weighted sum of its input and subsequently applies a compressor function $f : \mathbb{R} \rightarrow \mathbb{R}$ to produce its output activation. It can be shown that for every desired input-output mapping in the form of a real valued continuous function $\tilde{f}_d : \mathbf{x} \in \mathbb{R}^K \rightarrow \mathbb{R}$ and, for a non constant bounded and monotonically increasing activation function $f(\cdot)$ at all hidden elements, an integer M, an $M \times K$ matrix $\mathbf{U} = \left[u_{ij}\right]$ and M-dimensional vectors $\mathbf{v} = \left[v_j\right]$ and $\mathbf{b} = \left[b_j\right]$ exist such that

$$\max_{\mathbf{x} \in \Gamma} \left| \tilde{f}_d(\mathbf{x}) - \mathbf{v}^T f(\mathbf{U}\mathbf{x} - \mathbf{b}) \right| < \varepsilon \tag{16}$$

where ε is an arbitrarily small positive constant and Γ is a bounded close subset of $\mathbb{R}^K$. Note that even if it may be theoretically possible to find the network weights that make the output error as small as desired, in real situations the parameters' optimization is very hard due to the fact that in supervised learning the adjustment of parameters is generally based on a limited number of training pairs $\left(\mathbf{x}, \tilde{f}_d(\mathbf{x})\right)$. Moreover in noise filtering applications, the mapping of the noise signal to the corresponding clean signal is not usually a mathematical function $\tilde{f}_d(\cdot)$, and this violates one of the existence conditions of the above-stated theorem. For the filter adaptation, a backpropagation approach is usually used.

Neural network structures can also be successfully used in the transformed domain [28] to carry out the enhancement process, following a suitable training phase. The approach followed is generally based on a multistage architecture, comprising:

1. processing of the original data into a transform domain
2. nonlinear enhancement mapping performed by a neural network

The phase information is typically left unchanged through the overall process. From this perspective, if we give an estimate of noise power spectral density as input to the NN, we can see such a method as a form of nonlinear SS. This helps to address the nonlinear link between noise and speech due to the nature of transform that is not necessarily the one attainable through the Fourier transform (like log-power spectral, cepstral, LPC, and so on).

Several researchers have performed interesting studies on this subject. We can cite as examples the one employing time delay neural network for Mel-scaled spectral estimation [21] and the one with missing data technique using Reurrent Neural Networks [52]. Furthermore, the SS technique based on nonlinear spectral estimation [12] described above can also be interpreted within this framework.

2.4 Overview of Nonlinear Unsupervised Techniques

The problem of finding the maximum likelihood estimates of the speech and the model parameters, given the noisy data, has been successfully addressed by Wan and Nelson [16] [28] using neural autoregressive models and the Extended Kalman Filtering (EKF) method. The speech model in the time domain is the following non-linear autoregressive model:

$$\begin{aligned} s(k) &= f\left(s(k-1),\cdots,s(k-K),\mathbf{w}\right)+v(k) \\ y(k) &= s(k)+n(k) \end{aligned} \tag{17}$$

where $v(k)$ is the process noise in state equation, usually assumed to be white, and K is the model time length. A different model is used for each frame into which the noisy signal is segmented. The EKF method is able to yield the ML optimal estimate if the model is known. However, if no suitable data set for training is provided, the model parameters have to be learnt from the available observable sequence. Kalman Filter theory can be directly applied to the autoregressive model above, if we rewrite it in the state-space form and $f(.)$ is assumed to be linear:

$$\begin{aligned} \mathbf{s}(k) &= F\left[\mathbf{s}(k-1)\right]+Bv(k) \\ y(k) &= C\mathbf{s}(k)+n(k) \end{aligned} \tag{18}$$

where the following hold:

$$\begin{aligned} &\mathbf{s}(k)=\left[s(k),\cdots,s(k-K+1)\right]^T \\ &F\left[\mathbf{s}(k)\right]=\left[f\left(s(k),\cdots,s(k-K+1),\mathbf{w}\right),s(k),\cdots,s(k-K+2)\right]^T . \\ &C=[1\ 0\ \cdots\ 0] \qquad\qquad B=C^T \end{aligned} \tag{19}$$

The EKF algorithm is simply a generalization of the well-known KF when $f(.)$ is nonlinear, providing an approximation of $f(.)$ with a time-varying linear function. The EKF formulas are listed below, where $\sigma_v^2(k), \sigma_n^2(k)$ represent the variances of the process and observation noises respectively.

$$\begin{gathered}
\hat{\mathbf{s}}^-(k)=F\left[\hat{\mathbf{s}}(k-1),\hat{\mathbf{w}}(k-1)\right] \\
P_{\hat{\mathbf{s}}}^-(k)=AP_{\hat{\mathbf{s}}}(k-1)A^T+B\sigma_v^2(k)B^T \qquad A=\frac{\partial F\left[\hat{\mathbf{s}}(k-1),\hat{\mathbf{w}}\right]}{\partial\hat{\mathbf{s}}(k-1)} \\
G(k)=P_{\hat{\mathbf{s}}}^-(k)C^T\left(CP_{\hat{\mathbf{s}}}^-(k)C^T+\sigma_n^2(k)\right)^{-1} \\
P_{\hat{\mathbf{s}}}(k)=\left(I-G(k)C\right)P_{\hat{\mathbf{s}}}^-(k) \\
\hat{\mathbf{s}}(k)=\hat{\mathbf{s}}^-(k)+G(k)\left(y(k)-C\hat{\mathbf{s}}^-(k)\right)
\end{gathered} \tag{20}$$

However, note that one cannot exclusively rely on such a procedure to get what is required, i.e. a simultaneous estimation of the speech model and speech signal. As a result, a new set of state-space equations for neural networks weights $\mathbf{w}$ (used for nonlinearity parameterization) are formulated as follows:

$$\begin{aligned}
\mathbf{w}(k)&=\mathbf{w}(k-1)+\alpha(k) \\
y(k)&=f\left(\mathbf{s}(k-1),\mathbf{w}(k)\right)+v(k)+n(k)
\end{aligned}. \tag{21}$$

The neural system f(.) allows a nonlinear time-varying observation on $\mathbf{w}$. An EKF algorithm can be applied to yield an ML estimate of the current state assuming the other state $\mathbf{s}$ is known. The result is that we have two EKFs running in parallel (see Fig.3), one for state and the other for weights estimation. At each time step, the

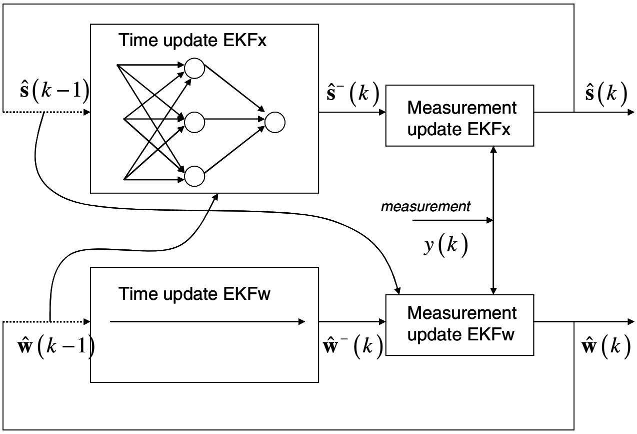

Fig. 3. The Dual Extended Kalman Filter method for speech enhancement

present state estimate $\hat{\mathbf{x}}(k)$ is used by the weight filter EKFw and the present weight estimate $\hat{\mathbf{w}}(k)$ feeds the state filter EKFx.

This approach, as is the case with SS, does not depend on the type of signal that we are dealing with and requires a suitable estimation of noise statistics. The main drawback is represented by the high computational cost occurring in training neural networks on-line, and some partial solutions have been proposed in order to reduce the computational complexity and obtain a faster convergence.

The Noise-Regularized Adaptive Filtering (NRAF) approach for speech enhancement [28] [31] involves a window based and iterative process that is similar to the dual EKF method, but does not use an AR model for the speech. It can be considered as a direct time-domain mapping filter (in the sense developed in [31]) avoiding the need for a clean dataset to train the network.

The objective of direct filtering approaches is to map the noisy vector **y**(k) to an estimate of the speech signal $\hat{s}(k) = f(\mathbf{y}(k))$. The neural network performing the mapping is trained by minimizing the mean-square error (MSE) cost function:

$$\min_f E\left\{\left[s(k) - f(\mathbf{y}(k))\right]^2\right\}. \tag{22}$$

We will now show how to minimize such a quantity without assuming that the clean signal s(k) is known. Consider the expansion:

$$\begin{gathered} E\left\{\left[s(k) - f(\mathbf{y}(k))\right]^2\right\} = E\left\{\left[y(k) - f(\mathbf{y}(k))\right]^2\right\} + 2E\left\{n(k) f(\mathbf{y}(k))\right\} + \\ -2E\left\{y(k) n(k)\right\} + E\left\{n^2(k)\right\} \end{gathered} \tag{23}$$

Since the last two terms are independent of *f*(.), it suffices to minimize the following alternative cost function to get the optimal solution:

$$\min_f \left\{E\left\{\left[y(k) - f(\mathbf{y}(k))\right]^2\right\} + 2E\left\{\left[n(k) f(\mathbf{y}(k))\right]^2\right\}\right\}. \tag{24}$$

A relevant advantage arises: the clean speech is not needed. Indeed the first term (corresponding to the cost associated with filtering the noisy signal itself) only depends on the observables, whereas the second term (namely the regularization term) on the noise statistics. An approximate solution is typically used for the latter. It is obtained by using the Unscented Transformation (UT), a method for calculating the statistics of a random variable going through a nonlinear transformation. Then, at each time instant, the network input is a suitable set of *K* vectors carrying the information related to the first and second order signal statistics, whereas the corresponding output is a weighted sample mean. A standard gradient based algorithm like back-propagation can be used to accomplish the minimization. The effectiveness of the method relies on the assumption that the accuracy of the second-order UT based approximation to the regularization term is good enough to achieve the network convergence to the true minimum MSE. Moreover, as in the Dual EKF approach, also in NRAF the speech non-stationarity can be dealt with by windowing the noisy data into short overlapping frames with a new filter for each frame.

In addition, Monte-Carlo simulation based approaches for audio signal enhancement have been recently proposed in some scientific works [4] [29-30]. Here their basic principles shall be discussed. Considering the clean and noisy speech as sequences of scalar random variables, we can assume they satisfy some kind of time-varying state-space equations, as previously done in (17). With a superior degree of generality, we can characterize our system by three deterministically nonlinear transition functions, here denoted as f, g, h. Function f is also dependent on discrete time k (therefore it will be represented as f_k), whereas the other two are in general dependent on the system parameter vector $\mathbf{w}(k)$ (and we will denote them as $g_{\mathbf{w}_k}, h_{\mathbf{w}_k}$ respectively). It follows that the state space equations are:

$$\begin{aligned} \mathbf{w}(k) &= f_k\left(\mathbf{w}(k-1), u(k)\right) \\ s(k) &= g_{\mathbf{w}_k}\left(s(k-1), v(k)\right) \\ y(k) &= h_{\mathbf{w}_k}\left(s(k), n(k)\right) \end{aligned} \tag{25}$$

where $u(k), v(k), n(k)$ are the innovation processes of the dynamical system, usually assumed to be statistically independent and identically distributed (i.i.d). As mentioned earlier, the involved functions in (25) are not linear, the parameter vector is not known (and possibly non-stationary) and the model is non-Gaussian. This results in severe computational difficulties in estimating the system parameters and/or the state signal (speech), and favors the usage of Monte Carlo simulations according to which the probability distributions are sampled and replaced by empirical distributions. It follows that the filtering and smoothing recursions occurring in state estimation can be simulated by means of the so called particle filters and smoothers developed through the point masses (particles) obtained from distribution sampling.

In [4], [30] and [31], speech signal process is modeled as a time-varying autoregressive (TVAR) model (as for the case of the Dual EFK method discussed above). Moreover the noise is assumed to be Gaussian and the coefficients of TVAR model a Gaussian random walk process. It can be shown that the following holds:

$$p\left(\mathbf{w}(k)\middle|y^{1:k}\right) \propto \int p\left(y(k)\middle|\mathbf{w}^{1:k}, y^{1:k-1}\right) p\left(\mathbf{w}(k)\middle|\mathbf{w}(k-1)\right) p\left(\mathbf{w}^{1:k-1}\middle|y^{1:k-1}\right) d\mathbf{w}^{1:k-1} \tag{26}$$

where $y^{1:k}$ stands for $\left\{y(1), \cdots, y(k)\right\}$, and accordingly for other variables occurring with same notation. The quantity $p\left(\mathbf{w}(k)\middle|y^{1:k}\right)$ is the filtering distribution, namely the objective of our estimation problem. Now, let us suppose to have an estimate of $p\left(\mathbf{w}^{1:k-1}\middle|y^{1:k-1}\right)$ at time instant $k-1$. This probability density function (pdf) can be sampled N times producing N different sample paths of $\mathbf{w}^{1:k-1}$, namely $\left\{\mathbf{w}_i^{1:k-1}, i=1,\cdots,N\right\}$. Hence, the particle approximation to $p\left(\mathbf{w}^{1:k-1}\middle|y^{1:k-1}\right)$ is given by:

$$p\left(\mathbf{w}^{1:k-1}\middle|y^{1:k-1}\right) \approx \sum_{i=1}^{N} \delta\left(\mathbf{w}^{1:k-1} - \mathbf{w}_i^{1:k-1}\right) \tag{27}$$

where $\delta(.)$ denotes the Dirac function. For each *i*, we can get the N samples $\left\{\mathbf{w}_i(k), i=1,\cdots,N\right\}$ from a proposal distribution $\pi\left(\mathbf{w}_i(k)\middle|\mathbf{w}_i^{1:k-1}, y^{1:k}\right)$, namely the importance distribution. Then we can use the latter samples to augment the former and generate the new sample paths at time instant *k*: $\left\{\mathbf{w}_i^{1:k}, i=1,\cdots,N\right\}$. A typical assumption is to set:

$$\pi\left(\mathbf{w}_i(k)\middle|\mathbf{w}_i^{1:k-1}, y^{1:k}\right) = p\left(\mathbf{w}(k)\middle|\mathbf{w}(k-1)\right) \tag{28}$$

It must be said that $p\left(\mathbf{w}(k)\middle|\mathbf{w}(k-1)\right)$ is fixed once we have chosen to apply a constrained Gaussian random walk in the TVAR coefficient domain.

Equation (27) can be substituted into (26), resulting in:

$$p\left(\mathbf{w}(k)\middle|y^{1:k}\right) \approx \sum_{i=1}^{N} \theta_i(k)\,\delta\left(\mathbf{w}^{1:k} - \mathbf{w}_i^{1:k}\right) \tag{29}$$

where $\theta_i(k)$ are the importance weights and $\theta_i(k) \propto p\left(y(k)\middle|\mathbf{w}_i^{1:k}, y^{1:k-1}\right)$ holds, as a direct consequence of (28). Under the assumption of conditionally linear Gaussian structure, the distribution $p\left(y(k)\middle|\mathbf{w}_i^{1:k}, y^{1:k-1}\right)$ can be evaluated efficiently using the Kalman filter and the prediction error decomposition. Indeed, our system model satisfies such a condition, as confirmed by (18). This ensures also $O(N)$ computational complexity and storage requirements for our algorithm. We have now an estimate of $p\left(\mathbf{w}(k)\middle|y^{1:k}\right)$ and we can iterate the procedure for all subsequent time instants. Furthermore, it follows that the MMSE estimate of the clean speech plus parameter vector of our system model ($f_{k|k}\left(\mathbf{s}(k), \mathbf{w}^k\right) = \left(\mathbf{s}(k), \mathbf{w}^k\right)$) is

$$\hat{I}_N\left(f_{k|k}\right) \triangleq \sum_{i=1}^{N} \tilde{\theta}_i^{0:k} E_{p\left(\mathbf{s}(k)\middle|\mathbf{w}^{1:k}, y^{1:k}\right)}\left[f_{k|k}\left(\mathbf{s}(k), \mathbf{w}_i^k\right)\right] \tag{30}$$

where $p\left(\mathbf{s}(k)\middle|\mathbf{w}^{1:k}, y^{1:k}\right)$ is a Gaussian distribution whose parameters may be computed using the Kalman filter and E is the expectation operator. According to the principle of Sequential Importance Sampling (SIS), satisfied by our choice of the proposal distribution, a recursive evaluation of the importance weights is allowed, which implies: $\theta(\mathbf{w}_{1:k}) = \theta(\mathbf{w}_{1:k-1})\theta(k)$. Finally, the normalized importance weights appearing in (30) are given by $\tilde{\theta}_i^{0:k} \triangleq \theta(\mathbf{w}_{1:k}^i) \Big/ \sum_{j=1}^{N} \theta(\mathbf{w}_{1:k}^j)$.

3 Binaural Nonlinear Noise Cancellation for Speech Enhancement

3.1 Review of Adaptive Noise Cancellation (ANC)

The classical scheme for the Adaptive Noise Cancellation (ANC) was originally proposed by Widrow et al. [32], and has been the subject of numerous studies involving a wide range of applications. In contrast to other enhancement techniques, no a priori knowledge of signal or noise is required for the method to be applied, but this advantage is paid for by the need of a secondary or *reference* input. This reference input should contain little or no signal but it should contain a noise measurement which is correlated, in some unknown way, with the noise component of the *primary* input. An important step in ANC is obtaining a reference signal which satisfies the above mentioned requirements. Referring to Fig.4, given a noisy speech (primary) signal $y[k]$, and assuming that $s[k]$ is uncorrelated with $n_1[k]$ and $n_2[k]$, and that $n_2[k]$ is processed by a linear filter $h[k]$ (generally non-causal), it is easy to show that $E\{e^2[k]\}$ is minimized when $v[k] = n_1[k]$, so that the output speech $e[k] = s[k]$ is the desired clean signal. Hence, the adaptive filter in classical linear methods is designed to minimize $E\{n_1[k] - v[k]\}$, using standard algorithms, like the least mean squares (LMS) technique.

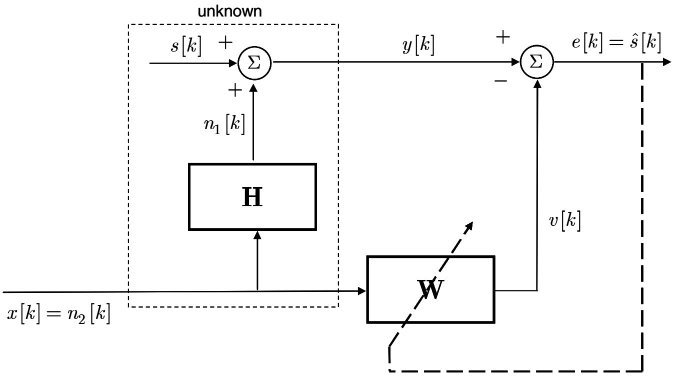

Fig. 4. Adaptive noise cancellation scheme for general nonlinear environments

3.2 Nonlinear ANC: Review of Approaches

Linear adaptive filtering, previously described, with the mean squared error (MSE) criterion is a standard signal processing method, and the reason for its success is the relative simplicity of design and ease of implementation. Nevertheless it can not often realize the Bayes conditional mean, which is the optimal filter for the MSE criterion, and generally a nonlinear function of the observed data. An important exception is if the observed data and the data to be estimated are jointly Gaussian: in this case the Bayes filter is a linear function. Since many real world signal processing applications have to deal with non-Gaussian signals, the use of a linear finite impulse response (FIR) or infinite impulse response (IIR) filter does not permit to obtain an acceptable level of noise or interference cancellation, because it can not efficiently approximate the nonlinear mapping between the known reference and the unknown interference signal. With reference to Fig.4, we say that the reference noise is related to the interference signal by an unknown nonlinear operator $\mathbf{H}$, approximated by a nonlinear feed-forward network. The objective is to determine the unknown nonlinear operator $\mathbf{H}$ by a nonlinear filter $\mathbf{W}$, so that we can optimally estimate the noise $n_1[k]$ and subtract it from the signal $y[k]$. In this way the primary source signal can be estimated. In the literature, a number of different techniques to design the filter $\mathbf{W}$ can be found which can be conveniently grouped in three principal classes: higher order-statistic filters, polynomial filters (in particular Volterra filters) and different kinds of neural networks. Higher order statistics (HOS) filters are based on ordering properties of input signals. A well-known member of this family is the Median filter, that is useful in removing impulsive noise, but poor in case of Gaussian noise. In [41] third-order statistics are used to derive novel design techniques which are more insensitive to corruption of the primary signal by additive Gaussian noise, compared to the second-order statistics ones. Referring to Fig.4, under the hypothesis that all signals are zero mean and stationary, and that $s[k]$ is independent of both $n_1[k]$ and $n_2[k]$ and that $n_1[k]$ and $n_2[k]$ are someway correlated, the optimal filter $\mathbf{W}$ can be determined using the third-order moment by solving the following

$$\sum_{i=0}^{q} w_3[i] R_{n_2}^{(3)}(m+i, l+i) = R_{yn_2}^{(3)}(m, l) \tag{31}$$

where

$$\begin{aligned} R_{n_2}^{(3)}(m,l) &\triangleq E\{n_2[k]n_2[k+n]n_2[k+l]\} \\ R_{yn_2}^{(3)}(m.l) &\triangleq E\{y[k]n_2[k+n]n_2[k+l]\} \end{aligned} \tag{32}$$

and different estimates can be obtained for different values of (m,l). Furthermore, if n_1 is linearly related to n_2 (i.e. if $\mathbf{H}$ in Fig.4 can be modeled by a linear time invariant (LTI) filter) then theoretically $w_3[k]$obtained from (31) is equivalent to those obtained by classical MSE methods, and it leads to complete cancellation of the interference, by identifying the true $\mathbf{H}$ filter. In practice, the theoretical auto- and cross-correlations are substituted by consistent sample estimator computed from the available data. The Volterra Filter (VF) has the important property to be linear in its parameters. So the identification of vector $\mathbf{H}$ in the MMSE sense can be obtained through the resolution of a linear equation. To find the optimal filters, we can operate both in the time and in the frequency domain [53]. An adaptive resolution of this equation is the RLS algorithm, which is based on the recursive calculation of the covariance matrix of the input signal of the filter. For the application of VF to the problem of noise cancellation, we refer to Fig.4 [42]. If $n_2[k]$ and $s[k]$ are independent and zero mean, then the previously mentioned algorithms can be used, if we replace $x[k]$ with $n_2[k]$ and $y[k]$ by $s[k]+n_1[k]$.

Next, a number of selected approaches to the interference cancellation problem using neural network filters are briefly analyzed, though other techniques can of course be found in the literature too.

As previously noticed, the problem of noise filtering can be viewed as the problem of finding the mapping of noisy signal patterns $y[k]$ to the corresponding noise-free signal patterns $s[k]$. According to this perspective, different kinds and topologies of neural networks can be used relating to the different relations between $n_1[k]$ and $n_2[k]$. Since a two layer feed-forward network has been proven capable of approximating any continuous non-linear mapping, assuming there are a sufficient number of hidden units, various implementations of this structure (with different contrast functions and number of hidden units) can be found in the literature. In [42], for example, a perceptron with one hidden layer and one output unit is used for the filter $\mathbf{W}$ (referencing to Fig.4 for the notation).

Denoting $\mathbf{n}_2[k]=\left[n_2[k],n_2[k-1],\cdots,n_2[k-K]\right]$, the mapping is described by

$$v[k]=\sum_{m=1}^{M}c_m \tanh\left(\mathbf{w}_m^T\mathbf{n}_2[k]-b_m\right) \tag{33}$$

where M is the number of hidden units, c_m and the vectors $\mathbf{w}_m$ are the weights coefficients, and b_m are the biases. The training is performed using the classical backpropagation technique. No method currently exists to precisely determine the optimal solution. Performance depends on the initial weights, the learning rate and the amount of training, but for small K from (33) the perceptron seems to perform a good approximation of the optimum Bayes filter.

The last kind of neural network analyzed in this work for the problem of noise cancellation is the Hyper Radial Basis Function (HRBF) neural network, following the approach described in [40]. The main idea is to consider the mapping $\mathbf{W}$ in Fig.4 to be approximated as the sum of various radial basis functions, each one with its own prior. Defining f_m, $m=1,\cdots,M$ as these functions, the function to minimize is:

$$L(n_1)=\sum_{k=1}^{K}\left(\sum_{m=1}^{M}f_m\left(n_2[k]\right)-n_1[k]\right)^2+\sum_{m(1)}^{M}\gamma_m\left\|P_m f_m\right\|^2 \tag{34}$$

where P_m are stabilizers in Tikhonov's stabilization theory and γ_m are regularization parameters (real and positive). The approximate solution of (34) is given by:

$$\tilde{n}_1 = \sum_{m=1}^{M}\sum_{j=1}^{K} w_j^m G_j^m\left(\bar{n}_2, \mathbf{q}_j^m\right) \tag{35}$$

where w_j^m are weight parameters and G_j^m are Green's functions. Choosing a set of stabilizers whose Green's functions are Gaussian, the HRBS neural network becomes formally equivalent to a two layer neural network the hidden layer of which realizes an adaptive nonlinear transformation (with adjustable weight and center parameters).

3.3 Multi-(sub)band Processing for Binaural Speech Enhancement

Some researchers have looked to the human hearing system as a source of engineering models to approach the enhancement problem, with some modelling the cochlea and others utilizing a model of the lateral inhibition effect. Two or more relatively closely spaced microphones have been used in an adaptive noise cancellation scheme [35], to identify a differential acoustic path transfer function during a noise only period in intermittent speech. The extension of this work, termed the Multi-Microphone Sub-band Adaptive (MMSBA) speech enhancement system, applies the method within a set of sub-bands provided by a filter bank. The filter bank can be implemented using various orthogonal transforms or by a parallel filter bank approach. The idea of employing multi-band processing for speech enhancement has also been considered in other contributions focusing on the spectral subtraction technique [54-55]. In the MMSBA approach [36-38], the sub-bands are distributed non-linearly according to a cochlear distribution, as in humans, following the Greenwood model [56]. The conventional MMSBA approach considerably improves the mean squared error (MSE) convergence rate of an adaptive multi-band LMS filter compared to both the conventional wideband time-domain and frequency domain LMS filters, as shown in [36-38]. It is assumed that the speaker is close enough to the microphones so that environmental acoustic effects on the speech are insignificant, that the noise signal at the microphones may be modelled as a point source modified by two different acoustic path transfer functions, and that an effective voice activity detector (VAD) is available. In practice, the MMSBA based speech-enhancement systems have been shown to give the important benefit of supporting adaptive diverse parallel processing in the sub-bands, namely Sub-band Processing (SBP), allowing signal features within the sub-bands, such as the noise power, the coherence between the in-band signals from multiple sensors and the convergence behaviour of an adaptive algorithm, to influence the subsequent processing within the respective frequency band. The SBP can be accomplished with no processing, intermittent coherent noise canceller, or incoherent noise canceller. In the conventional MMSBA approach, linear FIR filtering is performed within the SBP unit and the LMS algorithm is used to perform the adaptation. In the non-linear MMSBA, Volterra Filtering based SBP has been applied (together with the RLS algorithm), leading to a significant improvement of results, especially in real noisy environments. The Magnitude Squared Coherence (MSC) has been applied by [58] to noisy speech signals for noise reduction and also successfully employed as a VAD for the case of spatially uncorrelated noises. A modified MSC has been used for selecting an appropriate SBP option within the MMSBA system [36].

In the newly proposed modified MMSBA architecture [38], Wiener filtering (WF) operation has been applied in two different ways: at the output of each sub-band adaptive noise canceller, and at the global output of the original MMSBA scheme. The employment of such post-processing (WF) within the MMSBA allows to deal with residual incoherent noise components that may result from the application of conventional MMSBA schemes, similar to the approach adopted in [57]. In both the proposed architectures, the role of WF is to further mitigate the residual noise effects on the original signal to be recovered, following application of MMSBA noise-cancellation processing.

Finally, the MMSBA framework also allows incorporation of cross-band effects to mimic human lateral inhibition effects. One possibility seems to extend the recently reported promising work of Bahoura and Rouat [59], who have shown that non-linear masking of a time-space representation of speech can be used to achieve simulated noise suppression for the monaural case, by discarding or masking the undesired (noise) signals and retaining the desired (speech) signals. They have demonstrated that this non-linear masking can enhance single-sensor or monaurally recorded speech by performing non-linear filtering with adaptive thresholding (based on the Teager Energy operator Bahoura and Rouat [60]) on a time-frequency (multi-band) representation of the noisy signal. In [61] the MMSBA system with linear filtering and two different adaptive sub-band binaural structures have been compared in the noise reduction problem.

4 General Multi-channel Nonlinear Speech Enhancement

This section deals with those nonlinear techniques for enhancement of speech signals when more than one microphone is present, specifically when an *M*-element microphone array is available. Compared to the single-channel case discussed in Section 2, the multiple sensors allow suitable spatial filtering of the incoming signals thereby gaining a relatively enhanced capability of interference suppression. Two main categories of works can be identified in this area. One is based on the development of a nonlinear microphone array system, where both complementary beamforming and nonlinear SS are carried out to yield the final enhancement. The other approach deals with Log-Spectra estimation within different noise reduction frameworks.

Let us start from the former [44]. The goal here is to enhance the speech signal through a spatial spectral subtraction method by using a complementary beamformer. The presence of two complementary directivity patterns results in nonlinear SS processing that avoid use of a speech pause detector - which is normally employed in a typical SS scheme (see above). As depicted in Fig.5, the observed signals pass through two different weight vectors, then summed in order to produce primary and reference signals defined as:

$$\begin{aligned} \tilde{Y}^{(p)}(l,\omega) &= 2S_0(l,\omega) + \sum_{d\in\Omega}\left(\mathbf{ga}_d(l,\omega) - \mathbf{ha}_d(l,\omega)\right)\cdot N_d(l,\omega) \\ \tilde{Y}^{(r)}(l,\omega) &= \sum_{d\in\Omega}\left(\mathbf{ga}_d(l,\omega) - \mathbf{ha}_d(l,\omega)\right)\cdot N_d(l,\omega) \end{aligned} \qquad (36)$$

where S_0 is the speech signal coming from the look direction (so coinciding with S if we consider the model (2)), $\mathbf{g},\mathbf{h}$ are the *M*-element complementary weight vectors,

Ω is the set of directions relative to the different interfering signals approaching the beamformer, N_d is the noise signal corresponding to the d-th direction. The quantities $\mathbf{ga}_d, \mathbf{ha}_d$ describe the directivity patterns, and $\mathbf{a}_d$ is the steering vector:

$$\begin{aligned} &\mathbf{a}_d(l,\omega) = \left[a_{1,d}(l,\omega), a_{2,d}(l,\omega), \cdots, a_{M,d}(l,\omega)\right] \\ &a_{m,d}(l,\omega) = \exp\left(j\omega x_m \sin\left(\theta_d(l)\right)/c\right) \end{aligned} \tag{37}$$

where c is the sound velocity, θ_d the d-th direction of arrival, x_m the coordinate of the *m-th* element of the array. The term $\sin\left(\theta_d(l)\right)$ in (37) implies that the steering vector depends on the frame number l due to the non-stationary location of noise contributions (hence such a dependency can be neglected in the case of "static" noise). It can be easily proved that, under assumptions of complementary directivity patterns and uncorrelation of arriving signals, the reference signal can be subtracted from the primary to yield S_0 without any speech pause detector. In formulas:

$$\hat{S}(l,\omega) = \frac{1}{2}\left[\left|\tilde{Y}^{(p)}(l,\omega)\right|^2 - E\left[\left|\tilde{Y}^{(r)}(l,\omega)\right|^2\right]\right]^{1/2} \cdot \exp\left(j\phi(\omega)\right) \tag{38}$$

where $\hat{S}(l,\omega)$ is the estimated speech signal and $\phi(\omega)$ a suitable phase function, coming from a conventional beamformer (delay-and-sum, DS) in the above approach (see Fig.5). In order to avoid occurrence of over-subtraction, a better performing frame-by-frame SS rule has been used in [44].

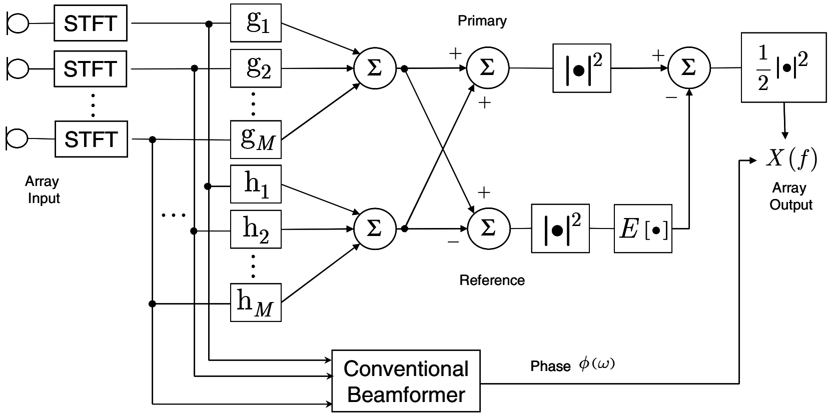

Fig. 5. Block diagram of the nonlinear mic array based on complementary beamforming

The directivity patterns are designed under the constraint of keeping the terms $\left|\mathbf{ga}_d(\omega)\cdot\mathbf{ha}_d(\omega)\right|$ as small as possible, for all ω, d, in order to have low noise contribution to the primary signal. This results in a nonlinear constrained least squares minimization problem, tackled by a suitable iterative procedure. The approach is supervised and the common choice made for the target directivity pattern involves setting the value 1 for the look direction and 0 otherwise. In such a way, it is possible to get lower sidelobes with respect to the DS array, resulting in a significant improvement of speech enhancement capability. However, the optimization procedure employed is not specifically oriented to minimize the average gain in each direction, causing a certain difficulty to reduce directional noise. That is why another optimization scheme, within the complementary beamforming based framework described above and depicted in Fig.5, has been proposed in [45]. According to this, the power spectrum of the estimated speech $\left|\hat{S}(l,\omega)\right|^2$ is calculated through a block averaging technique, giving origin to the quantity $\left|\hat{S}_B(\omega)\right|^2$ (where B is the number of blocks involved) that becomes the minimization criterion (assuming speech absence conditions). Again a relatively superior performance is obtained with respect to the conventional DS in situations where well-located noise sources (undesired speeches) are present. This occurs also when sound sources outnumber the microphones.

For the sake of completeness, we can mention the work of Dahl and Claesson [36], within the category of nonlinear microphone arrays. The approach followed in [46] is the one of nonlinear time-domain filtering, previously addressed in Section 2, and from this perspective it can be seen as a generalization of the single-channel approach described above. Hence no further details will be provided here.

Let us move now to address the Spectral Amplitude estimation based approach. The first contribution to consider here is the work of Lotter et al. [47] which provides two short-time spectral amplitude estimators generalizing the single-channel MMSE (Ephraim-Malah) and MAP [62] estimators. The method is based on the usual assumption that both speech and noise DFT coefficients have zero-mean equal-variance independent Gaussian pdfs. In the multi-channel case, the estimation of the speech spectral amplitudes is conditioned on complex spectra of M noisy channels $Y_m(\cdot)$, taking into account the notation used in (2):

$$\left|\hat{S}_m\right| = E\left\{\left|S_m\right| \left| Y_1, Y_2, \cdots, Y_M \right.\right\} \tag{39}$$

The above is calculated at each point of the time-frequency grid (p,ω). It can be showed that the new gain for channel m is:

$$\begin{aligned} G_{\bar{m}}(p,\omega) = \Gamma(1.5) \cdot \sqrt{\frac{R_{prio,\bar{m}}}{\left(1+R_{post,\bar{m}}\right)\left(1+\sum_{m=1}^{M} R_{prio,m}\right)}} \cdot \\ \cdot F_1\left(-0.5, 1, -\frac{\left|\sum_{m=1}^{M} \sqrt{\left(1+R_{post,m}\right) R_{prio,m}}\, e^{i\vartheta_m}\right|^2}{1+\sum_{m=1}^{M} R_{prio,m}}\right) \end{aligned} \tag{40}$$

where F_1 is the confluent hypergeometric series, Γ the Gamma function and ϑ_m the *m-th* noisy channel phase. Eq. (40) turns to (9) when $M = 1$, since (9) can be shown to be equal to:

$$G(p,\omega) = \Gamma(1.5)\cdot\sqrt{\frac{R_{prio}}{\left(1+R_{post}\right)\left(1+R_{prio}\right)}}\cdot F_1\left(-0.5,1,-\frac{\left(1+R_{post}\right)R_{prio}}{1+R_{prio}}\right) \tag{41}$$

It must be observed that (40) is obtained if perfect DOA (Direction of Arrival) correction is assumed within the microphone-array when the short-term spectral amplitude estimation $\left|\hat{S}(p,\omega)\right|$ is performed. As pointed out in [51], for DOA independent speech enhancement, the amplitude estimation has to be calculated by conditioning the expectation of the joint observation of noisy amplitudes, i.e. (39) turns to:

$$\left|\hat{S}_m\right| = E\left\{\left|S_m\right| \big| \left|Y_1\right|,\left|Y_2\right|,\cdots,\left|Y_M\right|\right\}. \tag{42}$$

In order to do the above in a simple and effective way, the authors in [47] suggested to employ the MAP estimator proposed originally for the single-channel case in [62]. It follows that, denoting $p(\cdot)$ as the probability density function (pdf) of a generic random variable, the following has to be maximized

$$\log(L) = \log\left(p\left(\left|Y_1\right|,\left|Y_2\right|,\cdots,\left|Y_M\right| \big| \left|S_m\right|\right)\cdot p\left(\left|S_m\right|\right)\right) \tag{43}$$

from which the following resulting gain can be derived:

$$\begin{aligned} G_{\bar{m}}(p,\omega) = {} & \frac{\sqrt{R_{prio,\bar{m}}\Big/\left(1+R_{post,\bar{m}}\right)}}{2\cdot\left(1+\sum_{m=1}^{M} R_{prio,m}\right)}\cdot \operatorname{Re}\left(\sum_{m=1}^{M}\sqrt{\left(1+R_{post,m}\right)R_{prio,m}} + \right. \\ & \left. + \sqrt{\left(\sum_{m=1}^{M}\sqrt{\left(1+R_{post,m}\right)R_{prio,m}}\right)^2 + (2-M)\left(1+\sum_{m=1}^{M} R_{prio,m}\right)}\right) \end{aligned} \tag{44}$$

which turns to the single-channel gain as follows:

$$G(p,\omega) = \frac{R_{prio} + \sqrt{R_{prio}{}^2 + \left(1+R_{prio}\right)R_{prio}\Big/\left(1+R_{post}\right)}}{2\cdot\left(1+R_{prio}\right)} \tag{45}$$

observing that the argument of $\operatorname{Re}(\cdot)$ is always a real number when $M = 1$.

Experimental results show how the new estimators allow a significant improvement of noise reduction performances (using segmental SNR as quality index) with respect to the single-channel EM rule in several operating conditions. Moreover as expected, the multi-channel MAP estimation approach turns out to be less sensitive to the phase errors (which are likely introduced by reverberation environments in rear-world applications) compared to the MMSE based method.

Along this direction we must cite the approach recently proposed by Cohen and Berdugo [49] that focused on the minimization of the Log-Spectra amplitude (LSA) distortion in environments where time-varying noise is present. The overall scheme (Fig.6) comprises an adaptive beamforming system (made of a fixed beamformer, a blocking matrix and a multi-channel adaptive noise canceller) and a suitable LSA estimation chain acting on the beamformer outputs, written as (in STFT domain):

$$\begin{aligned} V(l,\omega) &= S_1(l,\omega) + \tilde{N}_{1,st}(l,\omega) + \tilde{N}_{1,ns}(l,\omega) \\ U_m(l,\omega) &= S_m(l,\omega) + \tilde{N}_{m,st}(l,\omega) + \tilde{N}_{m,ns}(l,\omega) \qquad m = 1,\cdots,M \end{aligned} \tag{46}$$

where *st* and *ns* stand for stationary and non-stationary respectively. The objective is to find a suitable estimator of $S_1(l,\omega)$ minimizing the LSA distortion.

The noise cancellation system is responsible for reducing the stationary contribution and yielding the signal $V(l,\omega)$ on which the *optimally-modified log-spectral amplitude* (OM-LSA) gain function will be applied to achieve the goal. The evaluation of the nature of transient occurrences is performed through a suitable estimation of speech presence probability, which is based on a Gaussian statistical model and in particular on the transient beam-to-reference ratio (TBRR) defined as:

$$\Omega(l,\omega) = \frac{\mathcal{S}\left[V(l,\omega)\right] - \mathcal{M}\left[V(l,\omega)\right]}{\max\limits_{2\le m\le M}\left\{\mathcal{S}\left[U_m(l,\omega)\right] - \mathcal{M}\left[U_m(l,\omega)\right]\right\}} \tag{47}$$

where S[.], M[.] are the smoothing operator and the noise spectrum estimator arising by recursively averaging past spectral power values [48]. Assuming that the beamformer steering error is low and that the interfering noise is uncorrelated with speech, it can be said that a high TBRR means speech presence. When this is not the case, the noise estimation can be fast updated and then given to the OM-LSA [5] estimator for final speech enhancement. As confirmed by experimental results, such an approach seems to provide an adequate estimation of the time-varying noise spectral components and so a significant reduction of noise impact without degrading the speech

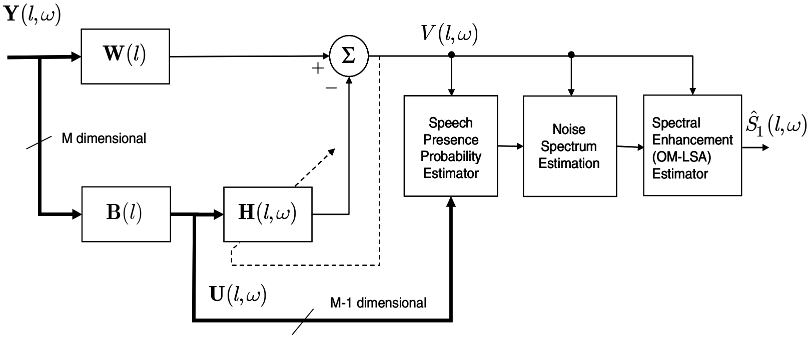

Fig. 6. Block diagram of the multi-microphone log-spectral amplitude estimation

information. Moreover, a significant improvement of performances is obtained through this multi-channel algorithm w.r.t. the single channel counterpart.

Another interesting approach based on Log-Spectra Amplitude estimation is the one employing a Supervised Regression technique, both in linear and nonlinear fashions. The method is termed Multiple Regression of the Log Spectrum (MRLS) [46] and has the objective of approximating the Log-Spectral Amplitude of the close-talking microphone (i.e. the original speech signal $s[k]$) by means of the Log-Spectral Amplitude of noisy signals emanating from other sensors. Mathematically, (taking (2) into account):

$$\log\left(S^{(d)}\right) \approx \sum_{m=1}^{M} \lambda_m \log\left(Y_m^{(d)}\right). \tag{48}$$

Let us first assume that the model (2) is written as:

$$y_m[k] = h_m[k] * s[k] + g_m[k] * n[k]. \tag{49}$$

Moving to the STFT domain, we can approximate the log-power spectrum of the m-th mic signal by a two-dimensional Taylor-series expansion around the reference Y_m^0 so that:

$$\log(Y_m) - \log\left(Y_m^0\right) \approx a_m\left(\log(S) - \log\left(S^0\right)\right) + b_m\left(\log(N) - \log\left(N^0\right)\right) \tag{50}$$

where it can be shown that the coefficients a_m, b_m depends on the SNR at m-th location. Now, considering $(\bullet)^{(d)}$ the deviation from $(\bullet)^{(0)}$, (50) turns to:

$$\log\left(Y_m^{(d)}\right) \approx a_m\left(\log\left(S^{(d)}\right)\right) + b_m\left(\log\left(N^{(d)}\right)\right). \tag{51}$$

The regression error is then given by the difference between the two terms in (48). The optimal weights λ_m can be obtained by minimizing such an error over a suitable number T of training samples, i.e.:

$$\varepsilon = \frac{1}{T}\sum_{t=1}^{T}\left\{\left[\log\left(S^{(d)}\right)\right]_t - \left[\sum_{m=1}^{M}\lambda_m \log\left(Y_m^{(d)}\right)\right]_t\right\}^2 . \tag{52}$$

The supervised optimization approach followed requires the employment of a close-talking microphone added to the available microphone array during the training phase the speech captured by the close talking mic is used as the speech signal S. The log-power spectrum is calculated though mel-filter bank (MFB) analysis and log operator. A cepstral based implementation has been also implemented, since the orthogonality of the Discrete Cosine Transform (DCT) transform ensures that minimization of (52) is equivalent to minimization in the cepstral domain. Several experimental results have shown that the MRLS approach allows a good approximation of the close-talking microphone and outperforms the adaptive beamformer from the perspective of speech recognition performance, ensuring also a low computational cost. Further improvements have been obtained when nonlinear regression (through Multi-Layer Perceptrons and Support Vector Machines) is employed [51]. A drawback is likely represented by the supervised optimization procedure that can be adopted within a speech recognition scheme, but turns out to be limiting in a more general framework for speech enhancement. An alternative approach to multi-channel non-linear speech enhancement has been described in [63], which applies neural network based sub-band processing (within the MMSBA processing framework) with promising initial results using real automobile reverberant data. This interesting approach warrants further investigation.

5 Concluding Summary

In this section, we summarize in tabular form for comparative purposes, the general features, list of operating assumptions, the relative advantages and drawbacks, and the various types of non-linear techniques for each class of speech enhancement strategy reviewed in this paper. Some references related to the methods not specifically described in the paper, are not included in the table.

5.1 Spectral Subtraction/Filtering Techniques

BASIC (LINEAR) TECHNIQUES [1], [2], [6]	
General Features	- based on MMSE estimator - optimal solution only for Gaussian statistics - frame by frame processing - linear subtraction/filtering in the noisy signal spectral domain
Assumptions	-speech and noise incorrelated - stationarity of noise signal - availability of a VAD system - zero mean Gaussian signals (for optimal estimation)
Advantages	- ease of implementation - computationally low demanding
Drawbacks	-non optimal estimate of clean speech - high level musical noise
NON LINEAR EXTENSIONS	
Ephraim Malah [8],[9], [10]	- nonlinear MMSE-STSA estimator - higher noise suppression compared to linear SS/SF - lower musical noise on the enhanced speech
Xie-Compernolle [12]	- empirical approach for the MMSE-STSA estimation - MLP to approximate the *Y*->*X* mapping - computationally more demanding (Monte Carlo simulations) - it requires the knowledge of the log spectral variance of the clean speech (supposed fixed)

5.2 Supervised NN Based Techniques

BASIC TECHNIQUES [21], [26], [52]	
General Features	- NN used to get the nonlinear mapping *Y->X* - off line training stage needed - standard training strategies (e.g. BP) to minimize MSE - different NN topologies can be used - iterative algorithms - filtering can be realized both in time and in different domains - speech parameters extracted form clean speech and used for enhancement
Assumptions	- additive Gaussian noise - both noise and speech stationarity - availability of clean speech data for off-line training - Gaussian signal statistics (for optimal estimation) - speech and noise stats representative of the training set - SNR or measures of joint noise-signal stats available
Advantages	- better estimation of clean speech - reduction of musical noise effect - good for fixed noise type
Drawbacks	- higher computational complexity - availability of a clean speech training set
OTHER EXTENSIONS	
Kalman Filter [3], [20]	- speech modeled as AR process - noise and speech variances available - no colored noise - involved parameters availability (noise gain, LPC parameters)
Switching Methods [16-19]	- a posteriori probability available for each class - suitable for non stationary signals - HMM estimate form noisy signal using Bayesian estimator form noisy speech - number of sates must be sufficient to model all ranges of signal and noise statistics
Extended KF [3], [20]	- non linear AR model for speech - ML convergence to AR parameters (off line training) - AR parameters available during enhancement - noise and speech variances must be known

5.3 Unsupervised Techniques

BASIC METHODS	
General Features	- NN used to get the nonlinear mapping *Y->x* - different topologies of NN can be used - iterative algorithm - joint estimation of signal and noise parameters - classical methods make use of EM technique
Assumptions	- noise less correlated than speech - short term stationarity of involved signals
Advantages	- no need of clean speech data
Drawbacks	- computationally demanding
OTHER EXTENTIONS	
Dual EKF [20], [26]	- ML estimates for both enhanced speech and parameters - can be used also with colored noise - usually used in the EM algorithm - high computational cost - frame by frame iteration

	- quasi stationarity for speech signal needed - possible initialization using HOS
NRAF [20], [28], [31]	- similar to dual EKF but no AR model for speech - time domain MMSE filtering (no clean speech needed)
Ephraim Cohen [4]	- speech modeled as TVAR system - Gaussian noise - lower computational complexity ($O(N)$)
Monte Carlo/Particle Filtering [4], [29], [30]	- computationally demanding

5.4 Adaptive Noise Cancellation (ANC)

ADAPTIVE NOISE CANCELLATION [31-33]	
General Features	- 2 channels available - no a priori knowledge on noisy signals required - reference channel contains no speech (ideally) - *s* and *n* incorrelated - linear or nonlinear filter can be used - ease of implementation and low computational cost (for linear filter) - using linear filter with MMSE estimator does not allow to get the Bayes conditional mean (optimal solution) - MMSE optimal filter: usually a nonlinear function of noisy data
NON-LINEAR ANC	
General Features	- non gaussian signal allowed - can deal with more complex mappings - higher computational complexity
HOS Filters [41]	- better to remove impulse noise - stationary zero mean signals - *s* and *n* independent
Volterra and NN [39-40], [42]	- different topologies allowed - training using classical algorithms
MMSBA [35-38]	- sub-band processing of noisy speech - different solution for the filter bank (DFT FB or Orthogonal transform) - sub-band distribution can be non linear - MMSE convergence improved - VAD available - different processing possible for different sub-bands - non linear filters can be used in the sub-bands processing

5.5 Multi-channel Speech Enhancement

MULTI CHANNEL TECHNIQUES	
General Features [43]	- array of *M* channel available - more degree of freedom - spatial filtering - enhanced capabilities of noise suppression
Complementary Beamforming and SS [44-45]	- spatial non linear SS technique - two complementary beamformers - no speech pause detector needed - incorrelation of arriving signals - nonlinear constrained least squares minimization - supervised approach - significant improvement of speech enhancement capability - optimized for directional noise in [44]

Dahl Cleasson [46]	- based on nonlinear time filtering - employment of supervised neural networks - generalization of the single channel case
Lotter [47]	- generalization of the single channel MMSE (Ephraim-Malah) - speech and noise DFT coefficients zero mean independent Gaussian pdfs - significant improvements of noise reduction performances - less sensitive to phase errors with MAP estimators
Cohen Berdugo [49]	- based on log-spectra amplitude (LSA) estimation - unsupervised approach - adaptive beamformer plus LSA estimation chain - noise cancellation system after the beamformer - suitable for environments with time-varying noises
Multiple Regression [50-51]	- supervised technique, both in linear and nonlinear fashion - based on LSA estimation of the speech signal, coming from the close-talking mic - implementation in the cepstral domain - suitable as front-end for speech recognition

References

1. Vaseghi, S.V.: Advanced Signal Processing and Digital Noise Reduction (2nd ed.). John Wiley & Sons, 2000
2. O'Shaughnessy, D.: Speech Communications – Human and Machine. IEEE Press, 2nd ed, Piscataway, NJ, 2000
3. Benesty, J, Makino, S., and Chen, J.,: Speech Enhancement. Signal and Communication Technology Series, Springer Verlag, 2005
4. Ephraim, Y., Cohen, I.: Recent Advancements in Speech Enhancement. The Electrical Engineering Handbook, CRC Press, 2005
5. Cohen, I. and Berdugo, B.H.: Speech enhancement for non-stationary noise environments. Signal Processing, vol. 81, pp. 2403-2418, 2001.
6. Boll S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech, Signal Process., ASSP-27:113-120, April 1979.
7. Lockwood, P., Boudy, J.: Experiment with a Nonlinear Spectral Subtractor (NSS). Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars. Speech Communications, 11, 215-228, 1992.
8. Ephraim, Y. Malah, D.:. Speech Enhancement Using a Minimum Mean Square Error Short Time Spectral Amplitude Estimator. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 1109-1121, 1984
9. Ephraim, Y. Malah, D.: Speech enhancement using a minimum mean square log spectral amplitude estimator. IEEE Trans. Acoust., Speech, Sig.Proc., vol 33, no 2, pp 443-445, 1985
10. Cappè, O.: Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech and Audio Proc., vol. 2, pp. 345 -349, April 1994
11. McAulay, R.J. and Malpass, M.L.: Speech Enhancement Using a Soft-Decision Noise Suppression Filter. IEEE Trans.on Acoust., Speech and Sig.Proc., vol. ASSP-28, no. 2, 1980
12. Xie, F. and Compernolle, D. V.: Speech enhancement by nonlinear spectral estimation - a unifying approach. EUROSPEECH'93, 617-620, 1993
13. Virag, N.: Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Processing, vol. 7, pp. 126–137, March 1999

14. Ephraim, Y. and Van Trees, H.L.: A signal subspace approach for speech enhancement. IEEE Trans. Speech and Audio Proc., vol. 3, pp. 251-266, July 1995
15. Lev-Ari, H. and Ephraim, Y.: Extension of the signal subspace speech enhancement approach to colored noise. IEEE Sig. Proc. Let., vol. 10, pp. 104-106, April 2003
16. Y. Ephraim: Statistical-model-based speech enhancement systems. Proc. IEEE, 80(10), October 1992
17. Ephraim, Y.: A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models. IEEE Trans. Signal Processing, vol. 40, pp. 725-735, Apr. 1992
18. Lee, K.Y., McLaughlin, S., and Shirai, K.: Speech enhancement based on extended Kalman filter and neural predictive hidden Markov model. IEEE Neural Networks for Signal Processing Workshop, pages 302-10, September 1996
19. Lee, J.; Seo; C., and Lee, K.Y. : A new nonlinear prediction model based on the recurrent neural predictive hidden Markov model for speech enhancement. ICASSP '02. vol. 1, pp.:1037-1040, May 2002
20. Wan, E.A., Nelson, A.T.: Networks for Speech Enhancement. Handbook of Neural Networks for Speech Processing, Edited by Shigeru Katagiri, Boston, USA. 1999
21. Dawson, M.I. and Sridharan, S.: Speech enhancement using time delay neural networks, Proceedings of the Fourth Australian International Conf. on Speech Science and Technology, pages 152-5, December 1992
22. Tamura, S.: An analysis of a noise reduction neural network. ICASSP ‘87, pp. 2001-4, 1987
23. Tamura, S.: Improvements to the noise reduction neural network, ICASSP ‘90, vol. 2, pp. 825-8, 1990
24. Knecht, W.G.: Nonlinear Noise Filtering and Beamforming Using the Perceptron and Its Volterra Approximation. IEEE Trans. On Speech and Audio Proc., vol.2, no.1, part 1, 1994
25. Knecht, W, Schenkel, M., Moschytz, G S.,: Neural Network Filters for Speech Enhancement. IEEE Trans. Speech & Audio Proc., 3(6),433-438, 1995
26. X-M. Gao, S.J. Ovaska, and I.O. Hartimo. Speech signal restoration using an optimal neural network structure, IJCNN 96, pages 1841-6, 1996
27. Gannot, S., Burshtein, D. and Weinstein, E.: Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms. IEEE Trans. Speech and Audio Proc., vol. 6, pp. 373-385, 1998
28. Wan, E.A., Nelson, A.T.: Neural dual extended Kalman filtering: applications in speech enhancement and monaural blind signal separation. Proceedings Neural Networks for Signal Processing Workshop, 1997
29. Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J.: Particle Methods for Bayesian Modeling and Enhancement of Speech Signals. IEEE Trans. Speech and Audio Processing, vol. 10, pp. 173 -185, Mar. 2002
30. Fong, W., Godsill, S.J., Doucet, A. and West, M.: Monte Carlo smoothing with application to audio signal enhancement. IEEE Trans. Signal Processing, vol. 50, pp. 438-449, 2002
31. Wan, E. and Van der Merwe, R.: Noise-Regularized Adaptive Filtering for Speech Enhancement. Proceedings of EUROSPEECH’99, Sep 1999
32. Widrow, B., Glover jr., J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Hearn, R. H., Zeidler, J. R., Dong jr., E. and Goodlin, R. C.: Adaptive Noise Cancelling: Principles and Applications. Proceedings of the IEEE, 63 (12): 1692–1716,1975
33. Clarkson, P.M.: Optimal and Adaptive Signal Processing. CRC Press, Boca Raton, 1993.
34. Toner, E.: Speech Enhancement Using Digital Signal Processing, PhD thesis, University of Paisley, UK, 1993

35. Darlington, D.J., Campbell, D.R.: Sub-band Adaptive Filtering Applied to Hearing Aids. Proc.ICSLP'96, pp. 921-924, Philadelphia, USA, 1996
36. 36.Hussain, A., Campbell, D.R.,: Intelligibility improvements using binaural diverse sub-band processing applied to speech corrupted with automobile noise. IEE Proceedings: Vision, Image and Signal Processing, Vol. 148, no.2, pp.127-132, 2001
37. Hussain, A., Campbell, D.R.: A Multi-Microphone Sub-Band Adaptive Speech Enhancement System Employing Diverse Sub-Band Processing. International Journal of Robotics & Automation, vol. 15, no. 2, pp. 78-84, 2000
38. Hussain, A., Squartni, S., Piazza, F.: Novel Subband Adaptive Systems Incorporating Wiener Filtering for Binaural Speech Enhancement. NOLISP05, ITRW on Non-Linear Speech Processing - LNAI 3817, Springer-Verlag, 2005.
39. Cha, I., Kassam, S.A.: Interference Cancellation Using Radial Basis Function Networks, . Signal Processing, vol.47, pp.247-268, 1995
40. Vorobyov, S.A., Cichocki, A.: Hyper Radial Basis Function Neural Networks for Interference Cancellation with Nonlinear Processing of Reference Signal. Digital Signal Processing, Academic Press, July 2001, vol. 11, no. 3, pp. 204-221(18)
41. Giannakis, G.B., Dandawate, A.V.: Linear and Non-Linear Adaptive Noise Cancellers. Proc ICASSP 1990. pp 1373-1376, Albuquerque, 1990
42. Amblard, P., Baudois, D.: Non-linear Noise Cancellation Using Volterra Filters, a Real Case Study. Nonlinear Digital Signal Processing, IEEE Winter Workshop on, Jan. 17-20, 1993
43. Brandstein, M.S. and Ward, D.B.: Microphone Arrays: Signal Processing Techniques and Applications. Springer-Verlag, Berlin, 2001
44. Saruwatari, H., Kajita, S., Takeda, K., Itakura, F.: Speech Enhancement Using Nonlinear Microphone Array Based on Complementary Beamforming, IEICE Trans. Fundamentals, vol.E82-A, no.8, pp.1501-1510, 1999.
45. Saruwatari, H., Kajita, S., Takeda, K., Itakura, F.: Speech Enhancement Based on Noise Adaptive Nonlinear Microphone Array, EUSIPCO 2000, X European Signal Processing Conference, Tampere Finland, 2000
46. Dahl, M. and Claesson, I.: A neural network trained microphone array system for noise reduction. IEEE Neural Networks for Signal Processing VI, pages 311-319, 1996
47. Lotter, T., Benien, C., Vary, P.: Multichannel Direction-Independent Speech Enhancement using Spectral Amplitude Estimation. Eurasip Journal on Applied Signal Processing, 11, pp. 1147-1156, 2003
48. I. Cohen, Berdugo, B.: Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement, IEEE Signal Processing Letters, vol.9, no.1 pp. 12-15, 2002
49. Cohen, I. and Berdugo, B.: Speech enhancement based on a microphone array and log-spectral amplitude estimation. Electrical and Electronics Engineers in Israel, the 22nd Convention, pp. 4.:6, Dec. 2002
50. Shinde, T., Takeda, K., Itakura, F.: Multiple regression of log-spectra for in-car speech recognition. ICSLP-2002, pp. 797-800, 2002
51. Li, W., Miyajima, C., Nishino, T., Itou, K., Takeda, K., Itakura, F.: Adaptive Nonlinear Regression using Multiple Distributed Microphones for In-Car Speech Recognition. IEICE Trans. Fundamentals, vol. E88-A, no. 7, pp. 1716-1723, 2005
52. Parveen, S. and Green, P.D.: Speech enhancement with missing data techniques using recurrent neural networks, Proc. IEEE ICASSP 2004, Montreal, 2004
53. Haykin, S. 2002 Adaptive Filter Theory (4th ed) Prentice Hall Information and System Science Series, Thomas Kailath Series Editor

54. Kamath, S. and Loizou, P.: A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, ICASSP 2002
55. Gülzow, T., Ludwig, L. and Heute, U.: Spectral-Substraction Speech Enhancement in Multirate Systems with and without Non-uniform and Adaptive Bandwidths. Signal Processing, vol. 83, pp. 1613-1631, 2003
56. Greenwood, V: A Cochlear Frequency-Position Function for Several Species-29 Years Later. J. Acoustic Soc. Amer., vol. 86, no. 6, pp. 2592-2605, 1990
57. Abutalebi, H. R., Sheikhzadeh, H., Brennan, R. L., Freeman, G.H.: A Hybrid Sub-Band System for Speech Enhancement in Diffuse Noise Fields, IEEE Sig. Process. Letters, 2003
58. Le Bouquin, R., Faucon, G.: Study of a Voice Activity Detector and its Influence on a Noise Reduction System. Speech Communication, vol. 16, pp. 245-254, 1995
59. Bahoura M. and Rouat J., “A new approach for wavelet speech enhancement”, Proc. EUROSPEECH, pp. 1937-2001, 2001
60. Bahoura M. and Rouat J., “Wavelet speech enhancement based on the Teager Energy Operator,” IEEE Signal Proc. Lett., 8(1), pp. 10-12, 2001
61. Cecchi, S, Bastari, A., Squartini, S. and Piazza, F.: Comparing Performances of Different Multiband Adaptive Architectures for Noise Reduction. Communications, Circuits and Systems (ICCCAS), 2006 International Conference of Guilin-China 2006
62. Wolfe, P.J. and. Godsill, S.J.: “Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement,” URASIP Journal on Applied Signal Processing, no. 10, pp. 1043–1051, 2003, special issue: Digital Audio for Multimedia Communications
63. Hussain, A., Campbell, D.R.: “Binaural sub-band adaptive speech enhancement using artificial neural networks,” Speech Communication, vol.25, pp.177-186, 1998, Special Issue: Robust Speech Recognition for Unknown Communication Channels