Functional Link Expansions for Nonlinear Modeling of Audio and Speech Signals

Danilo Comminiello, Simone Scardapane, Michele Scarpiniti, Raffaele Parisi, and Aurelio Uncini
Department of Information Engineering, Electronics and Telecommunications (DIET)
"Sapienza" University of Rome – Via Eudossiana 18, 00184 Rome, Italy
Email: danilo.comminiello@uniroma1.it

***Abstract*—Nonlinear distortions pose a serious problem for the quality preservation of audio and speech signals. In order to address this problem, such signals are processed by nonlinear models. *Functional link adaptive filter* (FLAF) is a linear-in-the-parameter nonlinear model, whose nonlinear transformation of the input is characterized by a basis function expansion, thus satisfying the universal approximation property. Since the expansion type affects the nonlinear modeling according to the nature of the input signal, in this paper we investigate the FLAF modeling performance involving the most popular functional expansions when audio and speech signals are processed. A comprehensive analysis is conducted in order to provide the best suitable solution for the processing of nonlinear audio signals. Experimental results are assessed also in terms of signal quality and intelligibility.**

I. Introduction

One of the most challenging problems to be addressed in the modeling of audio signals is the presence of nonlinear distortions. Nonlinearities may be introduced in audio signals by loudspeakers during large signal peaks, by vibrations of plastic loudspeaker shells, by D/A converters and power amplifiers, or also by some signal preprocessing. Such distortions affect audio and speech signals and impair their quality and intelligibility. This problem is particularly frequent in several digital audio applications involving the use of a loudspeaker, from high-quality listening to hands-free speech communications [1]–[3].

In order to compensate or reduce distortions affecting audio signals, nonlinear models are often adopted. In the literature, Volterra series expansions have been widely used for the modeling of loudspeaker distortions [4], [5]. Such solution has been successfully used for a large range of nonlinearities affecting audio and speech signals, but its high computational cost limits its use in real-time applications. Other kind of nonlinear transformation have been proposed to model different kinds of nonlinearities. Based on Fourier expansions, *even mirror Fourier nonlinear* (EMFN) filter is presented in [6], showing superior convergence properties than Volterra filters. In [7], a raised-cosine function is used to model both soft-clipping and hard-clipping nonlinearities. A flexible solution is proposed in [8], [9] based on spline functions, that are smooth parametric curves defined by interpolation of properly control points collected in a look-up table. *Functional link*-based models have been recently used for active noise control and nonlinear acoustic echo cancellation [3], [10].

In this paper, we focus on this last class of linear-in-the-parameter nonlinear model. The *functional link adaptive filter* (FLAF) [3] is characterized by a nonlinear expansion of the input signal and a subsequent linear adaptation in the higher dimensional space. The nonlinear transformation in the FLAF is performed by the *functional expansion block* (FEB), which contains the *functional links*, i.e. a series of linearly independent functions, which might be a subset of a complete set of orthonormal basis functions, satisfying universal approximation constraints. The use of functional links for the nonlinear expansion has been quite popular over the years for the identification of nonlinear systems, e.g. [10], [11], and recently it has been proposed also for echo cancellers [3], [12].

One of the main advantages of FLAF-based architectures lies in the flexibility, since the setting of several parameters is allowed in order to fit the model to a specific application. In this regard, an important choice in the FLAF design concerns the expansion type, i.e. the basis functions, or a subset of it, to be assigned for the functional link expansion. This choice mostly depends on the application and on the signals involved in the processing. Basis functions must satisfy universal approximation constraints and may be a subset of orthogonal polynomials, such as Chebyshev, Legendre, Laguerre and trigonometric polynomials, or just approximating functions, such as sigmoid and Gaussian functions. Among such functional expansions, the one based on *Chebyshev polynomials* has been widely used for several applications, in particular for nonlinear dynamic system identification [13] and channel equalization [14], showing a strong effectiveness. *Legendre expansion* was used as an alternative to Chebyshev polynomials in channel equalization [15] and, recently, even in real nonlinear system identification [16]. *Trigonometric polynomials* represent one of the most popular expansions [11], [17], besides being the only series to be used for applications involving audio and speech input signals [3], [10], although no other comparison with different expansion types is present in the literature. One of the first expansions used for functional links is the *random vector* (RV) [17]–[20]. Recently, similar models have been popularized under the name of *extreme learning machines* [21], [22], and used in the audio context when dealing with static music classification tasks [23], [24].

In this work, we want to provide a comprehensive analysis about the effect of different functional expansion types on the nonlinear modeling of distorted audio signals. Analyses are conducted by considering both error-based criteria and signal quality measures. Computational costs are also discussed. In particular, we assess the effects of different functional link expansions in some of the most representative digital audio applications involving nonlinear signals, such as the nonlinear

978-1-4799-1959-8/15/ $31.00 ©2015 IEEE

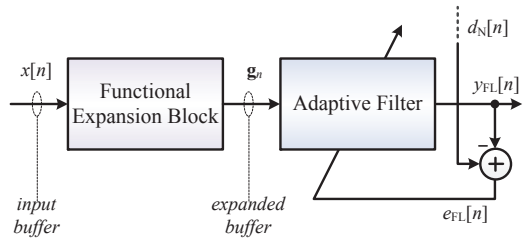

Fig. 1. The nonlinear functional link adaptive filter.

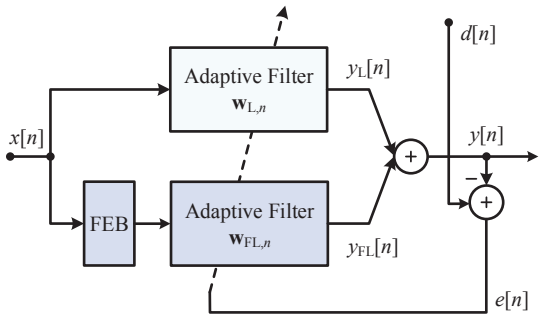

Fig. 2. The split functional link adaptive filter.

acoustic echo cancellation. This leads to determine the best choice for the functional link set, which can be considered as a benchmark for the processing of nonlinear audio signals.

The rest of the paper is organized as follows: Section II introduces the nonlinear FLAF model and the most popular functional expansions are described in Section III. After discussing a computational analysis in Section IV, we introduce the adopted performance evaluation criteria in Section V. Experimental results are shown in Section VI and in Section VII our conclusions are presented.

II. The Functional Link Approach for Nonlinear Modeling

In this section we present a brief overview of the FLAF model and of a FLAF-based architecture that we adopt to process audio input signals.

A. The Nonlinear FLAF Model

The *functional link adaptive filter* (FLAF) model is based on the representation of the input signal in a higher-dimensional space [17], in which an enhanced nonlinear modeling is allowed. Such approach derives from the machine learning theory, more precisely from the Cover's Theorem on the separability of patterns.

The FLAF is composed of two main parts: a nonlinear *functional expansion block* (FEB) and a subsequent linear filter, as depicted in Fig. 1. The FEB consists of a series of functions, which might be a subset of a complete set of orthonormal basis functions satisfying universal approximation constraints. The term "functional links" actually refers to the functions contained in the chosen set $\Phi = \{\varphi_0(\cdot), \varphi_1(\cdot), \ldots, \varphi_{Q-1}(\cdot)\}$, where Q is the number of functional links. At the n-th time instant, the FEB receives the input sample $x[n]$, which is stored in an input buffer $\mathbf{x}_{\mathrm{N},n} \in \mathbb{R}^{M_\mathrm{i}} = [\ x[n] \quad x[n-1] \quad \ldots \quad x[n-M_\mathrm{i}+1]\ ]^T$, where M_i is defined as the input buffer length. Each element of $\mathbf{x}_{\mathrm{N},n}$ is passed as argument to the chosen set of functions Φ, thus yielding a subvector $\overline{\mathbf{g}}_{i,n} \in \mathbb{R}^Q$:

$$\overline{\mathbf{g}}_{i,n} = \begin{bmatrix} \varphi_0(x[n-i]) \\ \varphi_1(x[n-i]) \\ \vdots \\ \varphi_{Q-1}(x[n-i]) \end{bmatrix}. \tag{1}$$

The concatenation of all the subvectors, for $i = 0, \ldots, M_\mathrm{i}-1$, engenders an *expanded buffer* $\mathbf{g}_n \in \mathbb{R}^{M_\mathrm{e}}$:

$$\begin{aligned} \mathbf{g}_n &= [\ \overline{\mathbf{g}}_{0,n}^T \quad \overline{\mathbf{g}}_{1,n}^T \quad \cdots \quad \overline{\mathbf{g}}_{M_\mathrm{i}-1,n}^T\ ]^T \\ &= [\ g_0[n] \quad g_1[n] \quad \ldots \quad g_{M_\mathrm{e}-1}[n]\ ]^T \end{aligned} \tag{2}$$

where $M_\mathrm{e} \geq M_\mathrm{i}$ represents the length of the expanded buffer. Note that $M_\mathrm{e} = M_\mathrm{i}$ only when $Q = 1$[1].

The achieved expanded buffer $\mathbf{g}_n$ is then fed into a linear adaptive filter $\mathbf{w}_{\mathrm{FL},n} \in \mathbb{R}^{M_\mathrm{e}} = [\ w_{\mathrm{FL},0}[n] \quad w_{\mathrm{FL},1}[n] \quad \ldots \quad w_{\mathrm{FL},M_\mathrm{e}-1}[n]]^T$, thus providing the nonlinear output:

$$y_{\mathrm{FL}}[n] = \mathbf{g}_n^T \mathbf{w}_{\mathrm{FL},n-1}. \tag{3}$$

Thereby, the nonlinear error signal is:

$$e_{\mathrm{FL}}[n] = d_{\mathrm{FL}}[n] - y_{\mathrm{FL}}[n] \tag{4}$$

which is used for the adaptation of $\mathbf{w}_{\mathrm{FL},n}$. In (4), $d_{\mathrm{FL},n}[n]$ represents the desired signal for the nonlinear model. Being $\mathbf{w}_{\mathrm{FL},n}$ a conventional linear filter, it can be adapted by any adaptive algorithm based on the minimization of the mean square error.

B. The Split FLAF Architecture

The nonlinear FLAF model is a purely nonlinear model, i.e., the adaptive filter receives as input only nonlinear elements. However, in audio and speech signal processing, very often there is also a linear component to adapt, as is the case of the acoustic impulse response in nonlinear acoustic echo cancellation. This is the reason why we adopt a FLAF-based structure that involves both linear and nonlinear filterings.

The *split functional link adaptive filter* (SFLAF) architecture [3], depicted in Fig. 2, is a parallel architecture including a linear path and a nonlinear path. The former is simply composed of a linear adaptive filter, which completely aims at modeling the linear components of an unknown system. Conversely, the nonlinear path is composed of a nonlinear FLAF. The SFLAF output signal results from the sum of the two path outputs:

$$y[n] = y_{\mathrm{L}}[n] + y_{\mathrm{FL}}[n] \tag{5}$$

[1]An exception is represented by the random vector expansion (see Subsection III-D) where, although $Q = 1$, we have $M_\mathrm{e} > M_\mathrm{i}$ due to the randomization process. In fact, the nonlinearity in (14) is not applied directly on the nonlinear input buffer $\mathbf{x}_{\mathrm{N},n}$, but on the randomized vector $\mathbf{z}_n$.

and, thereby, the overall error signal is:

$$\begin{aligned} e[n] &= d[n] - y[n] \\ &= d[n] - \mathbf{x}_{\mathrm{L},n}^T \mathbf{w}_{\mathrm{L},n-1} - \mathbf{g}_n^T \mathbf{w}_{\mathrm{FL},n-1} \end{aligned} \tag{6}$$

which is used for the adaptation of both the adaptive filters. In (6), $d[n]$ represents the desired signal including any near-end additive noise $v[n]$. The coefficient vectors $\mathbf{w}_{\mathrm{L},n}$ and $\mathbf{w}_{\mathrm{FL},n}$ may be adapted by any linear adaptive algorithm. In this paper we use a classic *normalized least-mean square* (NLMS) algorithm (see for example [25]):

$$\mathbf{w}_{\mathrm{L},n} = \mathbf{w}_{\mathrm{L},n-1} + \mu_{\mathrm{L}} \frac{\mathbf{x}_{\mathrm{L},n} e[n]}{\mathbf{x}_{\mathrm{L},n}^T \mathbf{x}_{\mathrm{L},n} + \delta} \tag{7}$$

$$\mathbf{w}_{\mathrm{FL},n} = \mathbf{w}_{\mathrm{FL},n-1} + \mu_{\mathrm{FL}} \frac{\mathbf{g}_n e[n]}{\mathbf{g}_n^T \mathbf{g}_n + \delta} \tag{8}$$

where δ is a regularization factor, and μ_{L} and μ_{FL} are the step-size parameters.

III. Functional Link Expansions

In the FLAF model described above, a fundamental role is played by the FEB, therefore, the choice of the set Φ of functional links represents a crucial decision. This choice may depend on the kind of application and on the kind of signals involved in the processing. The functional links must satisfy the universal approximation property, and they may be a subset of orthogonal polynomials or simple approximating functions. Here we introduce the most used functional link sets. Note that each expansion generates only nonlinear elements, since the processing of the linear elements is demanded to the linear path of the SFLAF.

A. Chebyshev polynomial expansion

Chebyshev polynomial functions are endowed with powerful nonlinear approximation capability. This is the reason why their use is widespread in different fields of application. In [13], [14], it was pointed out that an artificial neural network (ANN) with Chebyshev polynomial expansion has universal approximation capability and faster convergence than a multi-layer perceptron (MLP) network. The effectiveness of Chebyshev polynomial expansion is mainly due to the fact that it includes functions of previously computed functions. Moreover, Chebyshev expansion is based on power series expansion, which may approximate a nonlinear function with a very small error near the point of expansion. However, far from the point of expansion, the error may increase rapidly. With reference to other power series of the same degree, Chebyshev polynomials are quite computationally cheap and more efficient. However, when the power series converges slowly the computational cost dramatically increases. They were rarely used in the literature for the nonlinear transformation of audio signals (e.g. [26]).

Taking into account the i-th input sample $x[n-i]$ of the nonlinear input buffer $\mathbf{x}_{\mathrm{N},n}$, with $i = 0, \ldots, M_{\mathrm{i}}$, the Chebyshev polynomial expansion can be written as:

$$\begin{aligned} \varphi_j(x[n-i]) = 2x[n-i]\,\varphi_{j-1}(x[n-i]) \\ - \varphi_{j-2}(x[n-i]) \end{aligned} \tag{9}$$

for $j = 0, \ldots, P-1$, where P is the *expansion order*. Note that the number of functional links in a Chebyshev set is equal to the expansion order, i.e. $Q = P$. In (9), initial values (i.e., for $j = 0$) are:

$$\begin{aligned} \varphi_{-1}(x[n-i]) &= x[n-i] \\ \varphi_{-2}(x[n-i]) &= 1. \end{aligned} \tag{10}$$

B. Legendre polynomial expansion

Similar to Chebyshev polynomials, the Legendre functional links provide computational advantage while promising better performance [15]. Legendre polynomial expansions have been used for applications like channel equalization [15], in which Legendre-based *quadrature amplitude modulation* (QAM) equalizer performs better than radial basis function (RBF)-based and linear FIR-based equalizers. They were also recently used in audio applications [27].

Considering the i-th input sample, the Legendre polynomial expansion is described by:

$$\begin{aligned} \varphi_j(x[n-i]) = \frac{1}{j}\big\{(2j-1)\,x[n-i]\,\varphi_{j-1}(x[n-i]) \\ - (j-1)\,\varphi_{j-2}(x[n-i])\big\} \end{aligned} \tag{11}$$

for $j = 0, \ldots, P-1$. As for the Chebyshev expansion, the number of functional links is $Q = P$. Also in (11), initial values are set as (10).

C. Trigonometric series expansion

When trigonometric polynomials are used in upstream, i.e., before the adaptive filtering, the weight estimate will approximate the desired impulse response in terms of multidimensional Fourier series decomposition. In particular, compared with other orthogonal basis functions, trigonometric polynomials provide the best compact representation of any nonlinear function in the mean square sense, even for nonlinear dynamic systems as proved in [11]. Moreover, trigonometric functions are computationally cheaper than power series-based polynomials. Due to its properties, trigonometric series expansion is very popular in several applications [3], [10], [11], [17], including those on audio and speech signal processing.

Taking into account the i-th sample of the input buffer, it is possible to generalize the set of functional links using trigonometric basis expansion as:

$$\varphi_j(x[n-i]) = \begin{cases} \sin(p\pi x[n-i]), & j = 2p-2 \\ \cos(p\pi x[n-i]), & j = 2p-1 \end{cases} \tag{12}$$

where $j = 0, \ldots, Q-1$ is the functional link index and $p = 1, \ldots, P$ is the expansion index, being P the *expansion order*. In this case, the functional link set is composed of $Q = 2P$ functional links. Note that the expansion order for the trigonometric series is different from the order of both Chebyshev and Legendre polynomials. Equation (12) describes a memoryless expansion. A trigonometric expansion with memory can be easily achieved by adding the cross-product terms, as detailed in [3]. It may be very useful in the modeling of dynamic nonlinearities. In a trigonometric expansion with memory, the *memory order* K determines

the length of the additional functional links, i.e., the depth of the outer products between the i-th input sample and the functional links related to the previous input samples. Convergence analysis of trigonometric FLAF is shown in [28].

D. Random vector expansion

The last functional expansion type that we consider is the random vector (RV) functional link [18], [19]. The RV expansion is parametric with respect to a set of internal weights, which are randomly assigned at the beginning of the learning process. In particular, the expansion is performed in two steps. First, the nonlinear buffer at the n-th time instant $\mathbf{x}_{\mathrm{N},n}$ is fed into a randomization process, thus resulting in an expanded vector $\mathbf{z}_n \in \mathbb{R}^{M_\mathrm{e}}$:

$$\mathbf{z}_n = \mathbf{V}\mathbf{x}_{\mathrm{N},n} + \mathbf{b} \tag{13}$$

where the elements of $\mathbf{V} \in \mathbb{R}^{(M_\mathrm{e} \times M_\mathrm{i})}$ and $\mathbf{b} \in \mathbb{R}^{M_\mathrm{e}}$ are assigned from the uniform probability distribution in $[-1, +1]$. A sigmoid nonlinearity is then applied to the vector $\mathbf{z}_n \in \mathbb{R}^{M_\mathrm{e}}$, thus achieving:

$$\varphi\left(z\left[n-i\right]\right) = \frac{1}{1+e^{(-z[n-i])}}. \tag{14}$$

In the RV expansion, the set of functional links Φ is characterized by the only sigmoid function (i.e., $Q = 1$), which is applied to all the elements of the expanded vector $\mathbf{z}_n$. This is the reason why the functional link index j in (14) is omitted. It is worth noting that the sigmoid function is just one of the possible choice to apply a nonlinearity to the vector $\mathbf{z}_n$. Unlike the previous expansion types, the RV does not involve any expansion order, nor any memory. Therefore, the length M_e of the expanded buffer $\mathbf{g}_n$ does not depend on any other parameters and is *a priori* fixed. Convergence properties of the RVFL model are analyzed in [19].

IV. Computational Analysis of Functional Link Expansions

Here we want to compare the different functional link expansions introduced in the previous section in term of the computational cost demanded by the expansion. Therefore, we do not consider additional cost of any structure (e.g., the nonlinear FLAF or the SFLAF, or any other FL-based model) but we focus only on the operations made by the FEB.

The Chebyshev polynomial expansion is described by eq. (9), which requires $2PM_\mathrm{i}$ multiplications and PM_i additions at each iteration. The Legendre polynomial expansion is defined by eq. (11), which is a little bit more complex than the Chebyshev expansion. This reflects also an increase of the computational cost that overall involves $2P\left(2M_\mathrm{i}+1\right)$ multiplications and $P\left(M_\mathrm{i}+2\right)$ additions. The computational cost of the trigonometric series expansion has been extensively analyzed in [3], [29]. A memoryless trigonometric expansion, defined by (12), requires $P\left(M_\mathrm{i}+1\right)$ multiplications and M_e function evaluations (i.e., sines and cosines). Such function evaluations can be efficiently implemented by using lookup tables. Although this cost is cheaper with respect to the previous one, it may increase in case of a trigonometric expansion with memory. In fact, if we consider a memory order $K > 0$, we have an additional cost of $\left(M_\mathrm{i}-K\right)KP + \sum_{k=1}^{K-1} kP$ multiplications. The last expansion type is the random vector one, described by eqs. (13) and (14). For each iteration, the RV expansion requires $M_\mathrm{e}\left(M_\mathrm{i}+1\right)$ multiplications, $M_\mathrm{e}\left(M_\mathrm{i}+1\right)$ additions and M_e function evaluations (i.e., exponentials), thus representing the higher cost among those considered. Note that initializations of the $\mathbf{V}$ and $\mathbf{b}$ in (13) require an additional cost of $M_\mathrm{e}\left(M_\mathrm{i}+1\right)$ multiplications and $M_\mathrm{e}\left(M_\mathrm{i}+1\right)$ additions. However, since it is required only once, this additional cost is negligible with respect to the run-time computational cost.

TABLE I. Computational cost comparison of different functional link expansions in terms of multiplications.

Expansion Type	No. Multiplications
Chebyshev Polynomial Expansion	$2M_\mathrm{e}$
Legendre Polynomial Expansion	$4M_\mathrm{e}$
Trigonometric Series Expansion	$M_\mathrm{e}/2 + P$
Random Vector Expansion	$M_\mathrm{e}\left(M_\mathrm{i}+1\right)$

In order to have a clearer comparison between the different costs, we can express all of them in of M_e. In fact, we can consider that $M_\mathrm{e} = PM_\mathrm{i}$ for Chebyshev and Legendre polynomial expansions and $M_\mathrm{e} = 2PM_\mathrm{i}$ for trigonometric expansion. Therefore, we can think to fix M_e *a priori* and compare the four types of expansions. Results are collected (in terms of multiplications only) in Table I, where it is worth noting that, fixing *a priori* the expanded buffer length M_e and considering that usually $P << M_\mathrm{e}$, the trigonometric expansion implies the smallest number of multiplications, while the highest cost is drawn by the random vector expansion.

V. Performance Evaluation Criteria

The goal of this paper is to assess different functional link expansions within the modeling of speech and audio signals. This is the reason why we measure performance taking into account those quality indices that are thought for the evaluation of audio and speech signals.

The first important quality measure that we consider is the *echo return loss enhancement* (ERLE), which is the most significant index that is used in acoustic echo cancellation (AEC) applications. As depicted in Fig. 3, at the n-th time instant, the far-end signal $x\left[n\right]$ is fed through a loudspeaker-enclosure-microphone system (LEMS) activating an echo path. The loudspeaker may introduce an amount of distortion to the input signal, thus requiring a nonlinear modeling of $x\left[n\right]$. The microphone captures a reverberated and possibly distorted version of the input signal $s\left[n\right]$, that is the echo signal to be cancelled, and a near-end signal that we suppose consisting only of background noise $v\left[n\right]$. The overall microphone signal is also known as the desired signal $d\left[n\right]$. The SFLAF receives the reference signal $x\left[n\right]$, creates an estimate $y\left[n\right]$ of the signal passed through the LEMS and removes it from the desired signal $d\left[n\right]$, thus yielding the error signal $e\left[n\right]$ to be delivered to far-end. In this context, the ERLE measures the amount of echo signal, in decibels, which is canceled from the microphone signal, and it is expressed as:

$$\mathrm{ERLE}\left[n\right] = 10\log_{10}\left(\frac{\mathrm{E}\left\{d^2\left[n\right]\right\}}{\mathrm{E}\left\{e^2\left[n\right]\right\}}\right) \tag{15}$$

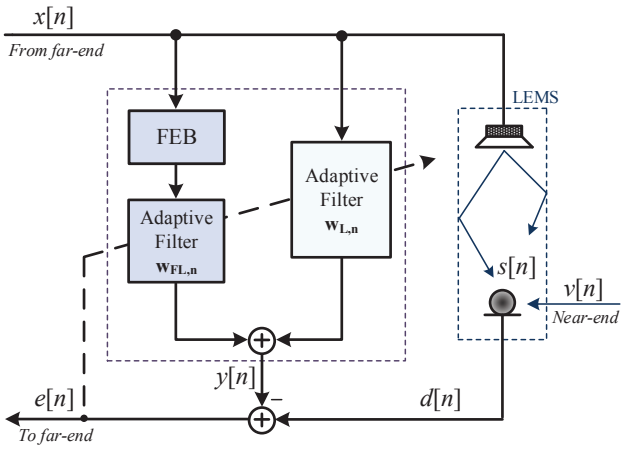

Fig. 3. Nonlinear acoustic echo canceller including the SFLAF.

where $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation. The ERLE is a fundamental measure denoting how much echo signal is canceled, but it does not reflect always a real enhancement of quality and intelligibility of the signal. Therefore, in support of the ERLE, we take into account other quality and intelligibility indices that are suitably designed for audio and speech signals. We use such measures to evaluate how "well" (from a quality point of view) the SFLAF model produces an estimate $y[n]$ of the echo signal $s[n]$.

In this context, one of the most important indices for the quality evaluation of a speech signal is the *perceptual evaluation of speech quality* (PESQ) objective measure [30], [31], which assesses speech quality by estimating the overall loudness difference between the original signal and its estimation. This measure has been widely adopted to predict the speech quality in enhancement processes [32]. The original and estimated signals, respectively $s[n]$ and $y[n]$, are first equalized to a standard listening level and filtered by a filter with response similar to that of a standard telephone handset. The signals are processed through an auditory transform to obtain the loudness spectra. The difference in loudness between the $s[n]$ and $y[n]$ is computed and averaged over time and frequency to produce the prediction of subjective quality rating [32]. The PESQ produces a score between 1.0 and 4.5, with high values indicating better quality.

While PESQ evaluates the speech quality in terms of loudness difference, the models based on linear predictive coding (LPC) look at the signal differences in the spectral envelopes. For the experimental results we have considered three different LPC-based measures: the *log likelihood ratio* (LLR), the *Itakura-Saito distance*, and the *cepstrum distance*. However, since experiments yielded the same indications for the three measures we focus only on the LLR that we can assume representative of all the three LPC-based measure. The LLR measure is defined as [33]:

$$\mathrm{LLR} = \log\left(\frac{\mathbf{a}_y^T \mathbf{R}_s \mathbf{a}_y}{\mathbf{a}_s^T \mathbf{R}_s \mathbf{a}_s}\right) \tag{16}$$

where $\mathbf{a}_s$ and $\mathbf{a}_y$ are the LPC vectors of $s[n]$ and $y[n]$, respectively, and $\mathbf{R}_s$ is the autocorrelation matrix of $s[n]$. The spectral envelope difference is as minimum as the LLR tends to zero. Signals are segmented in frames using Hamming windows of 30-ms duration with 75% overlap between adjacent frames [32]. Moreover, the LPC analysis is performed with order 10.

Another important objective measure is the *weighted spectral slope* (WSS) [34]. The WSS measures the weighted difference in each frequency band between the spectral slopes, which are obtained as the difference between adjacent spectral magnitudes in decibels. The WSS measure evaluated in this paper is defined as [32]:

$$\mathrm{WSS} = \frac{1}{N_\mathrm{f}} \sum_{i=0}^{N_\mathrm{f}-1} \frac{\sum_{j=1}^{N_\mathrm{b}} h_j[i] \left(S_{s,j}[i] - S_{y,j}[i]\right)^2}{\sum_{j=1}^{N_\mathrm{b}} h_j[i]} \tag{17}$$

where N_f is number of data frame, N_b is the number of critical bands (25 is the value suggested in [34]), $h_j[i]$ is a weight for the i-th frame and the j-th frequency band, and $S_{s,j}[i]$ and $S_{y,j}[i]$ are the spectral slopes for $s[n]$ and $y[n]$, respectively. As for the LLR, the signals are segmented in frames.

As regards the intelligibility of the processed signals, we focus on two indices, which are the *coherence speech intelligibility index* (CSII) and the *normalized covariance metric* (NCM). The CSII is an extension of the magnitude square coherence function suited for assessing the effects of distortions (e.g., peak clipping) on speech intelligibility by normal-hearing and hearing-impaired subjects [35]. The CSII is similar to the speech intelligibility index (SII), but it considers the signal-to-distortion ratio instead of the signal-to-noise ratio (SNR) (see [35] for a complete description). The CSII is composed of three measures that are computed by first dividing the segments into three level regions and computing separately the CSII measure for each region. A linear combination of the three values followed by a logistic function transformation is subsequently used to model the CSII intelligibility score (also named I3 in [35]). The other intelligibility index is the NCM [36], in which transmission index (TI) values are determined from the envelopes of the original and the estimated signals in each frequency band. Unlike the traditional speech TI (STI) measure, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function (MTF), the NCM measure is based on the covariance between the original and the estimated envelope signals [32]. A complete description of the NCM can be found in [36].

VI. Experimental Results

In this section, we assess the performance of the SFLAFs involving different functional expansions. We consider two kinds of input signals: the first one is a purely speech signal (from a male speaker), while the second one is a more general audio signal recorded from a radio station, which includes both speech and music and shows different volume levels. At each input signal we apply two different nonlinearities: a symmetrical soft-clipping and an asymmetric sigmoidal nonlinerity.

The symmetrical soft-clipping nonlinearity aims at simulating the classic saturation effect of a loudspeaker, and it affects the signal according also to its volume level. The soft-clipping nonlinearity that we use is described by the following

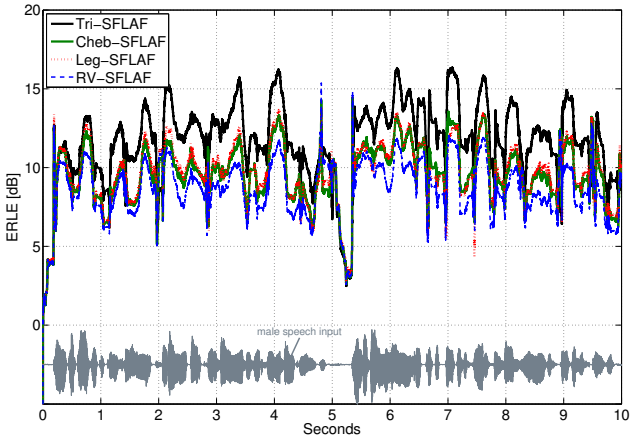

Fig. 4. Performance comparison in terms of ERLE between SFLAFs with different expansion types in case of speech input affected by a soft-clipping nonlinearity.

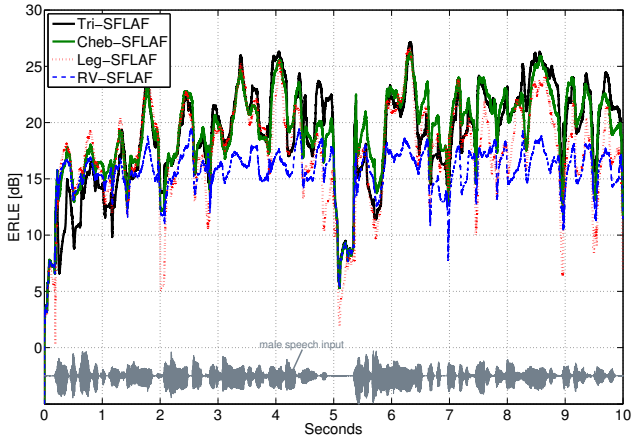

Fig. 5. Performance comparison in terms of objective measures between SFLAFs with different expansion types in case of speech input affected by an asymmetric sigmoidal nonlinearity using $M_{\mathrm{i}} = M$.

TABLE II. PERFORMANCE COMPARISON IN TERMS OF OBJECTIVE MEASURES BETWEEN SFLAFS WITH DIFFERENT EXPANSION TYPES IN CASE OF SPEECH INPUT AFFECTED BY A SOFT-CLIPPING NONLINEARITY.

Measure	Tri-SFLAF	Che-SFLAF	Leg-SFLAF	RV-SFLAF
PESQ	**3.865**	3.724	3.773	3.779
LLR	**0.380**	0.471	0.482	0.461
WSS	10.45	12.00	11.66	**10.35**
CSII	**0.687**	0.623	0.644	0.618
NCM	**0.856**	0.819	0.830	0.820

TABLE III. PERFORMANCE COMPARISON IN TERMS OF OBJECTIVE MEASURES BETWEEN SFLAFS IN CASE OF SPEECH INPUT AFFECTED BY AN ASYMMETRIC SIGMOIDAL NONLINEARITY FOR $M_{\mathrm{i}} = M$.

Measure	Tri-SFLAF	Che-SFLAF	Leg-SFLAF	RV-SFLAF
PESQ	3.911	3.826	**3.922**	3.623
LLR	**0.199**	0.317	0.239	0.307
WSS	**9.86**	12.19	11.59	14.22
CSII	**0.801**	0.761	0.790	0.669
NCM	0.884	0.878	**0.907**	0.843

expression [29]:

$$\overline{x}\left[n\right] = \begin{cases} \frac{2}{3\zeta}x\left[n\right] & \text{for} \quad 0 \leq \left|x\left[n\right]\right| \leq \zeta \\ \operatorname{sign}\left(x\left[n\right]\right)\frac{3-(2-|x[n]|/\zeta)^2}{3} & \text{for} \quad \zeta \leq \left|x\left[n\right]\right| \leq 2\zeta \\ \operatorname{sign}\left(x\left[n\right]\right) & \text{for} \quad 2\zeta \leq \left|x\left[n\right]\right| \leq 1 \end{cases} \tag{18}$$

where $0 < \zeta \leq 0.5$ is a nonlinearity threshold.

A second nonlinearity that we adopt is a memoryless asymmetric sigmoid function that aims at simulating the distortion produced by the coil movement of a loudspeaker. Such nonlinearity is introduced in [3] and it is described by:

$$\overline{x}\left[n\right] = \gamma\left(\frac{1}{1+e^{(-\rho q[n])}} - \frac{1}{2}\right) \tag{19}$$

where:

$$q\left[n\right] = \frac{3}{2}x\left[n\right] - \frac{3}{10}x^2\left[n\right]. \tag{20}$$

In (19), the parameter γ is the sigmoid gain and it is set equal to $\gamma = 2$, while ρ represents the sigmoid slope and it is chosen as:

$$\rho = \begin{cases} 4, & q\left[n\right] > 0 \\ \frac{1}{2}, & q\left[n\right] \leq 0 \end{cases}. \tag{21}$$

In both the cases of clipping (eq. (18)) and sigmoid (eq. (19)) nonlinearities, the distorted signal $\overline{x}\left[n\right]$ is convolved by a simulated acoustic impulse response measured at 8 kHz sampling rate and truncated after $M = 300$ samples. The resulting reverberated signal $s\left[n\right]$ is captured by a microphone along with the near-end background noise $v\left[n\right]$, which is additive Gaussian noise providing 20 dB of SNR. The length of each experiment is 10 seconds.

We use the following parameter setting for all the sets of experiments: step-size parameters $\mu_{\mathrm{L}} = 0.2$ and $\mu_{\mathrm{FL}} = 0.1$, regularization factor $\delta = 10^{-3}$, expansion order $P = 10$, and memory order $K = 0$ for the trigonometric expansion. The input buffer length M_{i} is set at each experiment. However, for all the experiments, we fix *a priori* the expanded buffer length $M_{\mathrm{e}} = PM_{\mathrm{i}}$ in order to yield a fair comparison in terms of nonlinear elements. MATLAB code for the SFLAFs is based on the `flaf` package freely available on BitBucket[2].

A. Male speech input

The first set of experiments is characterized by a male speech signal as input, on which we apply the two nonlinearities introduced above. First, we consider the soft-clipping nonlinearity (18), for which we use the same input buffer length $M_{\mathrm{i}} = 50$ (and consequently $M_{\mathrm{e}} = 500$) for all the expansions. Results in terms of the ERLE are depicted in Fig. 4, where it is rather clear that the SFLAF using the trigonometric series expansion achieves the best performance. We also evaluate the performance in terms of objective measures, collected in Table II, that also denotes a superiority of the trigonometric SFLAF both in quality and intelligibility.

We also consider the asymmetric sigmoidal nonlinearity (19), for which we choose a full input buffer length $M_{\mathrm{i}} = M$,

[2]https://bitbucket.org/ispamm/flaf/

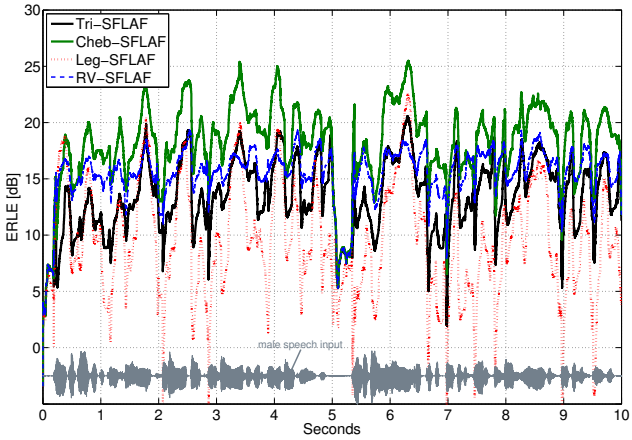

Fig. 6. Performance comparison in terms of objective measures between SFLAFs with different expansion types in case of speech input affected by an asymmetric sigmoidal nonlinearity using $M_{\mathrm{i}} = 20$.

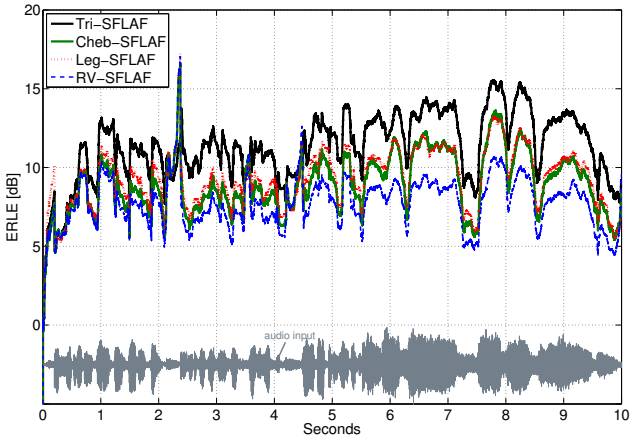

Fig. 7. Performance comparison in terms of ERLE between SFLAFs with different expansion types in case of audio input affected by a soft-clipping nonlinearity.

TABLE IV. PERFORMANCE COMPARISON IN TERMS OF OBJECTIVE MEASURES BETWEEN SFLAFS WITH DIFFERENT EXPANSION TYPES IN CASE OF AUDIO INPUT AFFECTED BY A SOFT-CLIPPING NONLINEARITY.

Measure	Tri-SFLAF	Che-SFLAF	Leg-SFLAF	RV-SFLAF
LLR	**0.074**	0.221	0.169	0.290
WSS	**6.50**	11.26	10.37	12.47
NCM	**0.993**	0.975	0.982	0.942

due to the strong distortion introduced. As it is possible to see in Fig. 5, except from the RV-SFLAF that shows the worst performance, the other SFLAFs achieve similar results in terms of the ERLE. However, from the objective measures in Table III, a slight advantage is provided by the trigonometric SFLAF and sometimes by the Legendre SFLAF. However, among several experimental results, it is also worth noting that if we choose a very small input buffer length (e.g., $M_{\mathrm{i}} = 20$), the trigonometric and Legendre SFLAFs degrades their performance, while the other two show also a slight decrease. Overall in this case, as depicted in Fig. 6, the Chebyshev SFLAF achieves the best result.

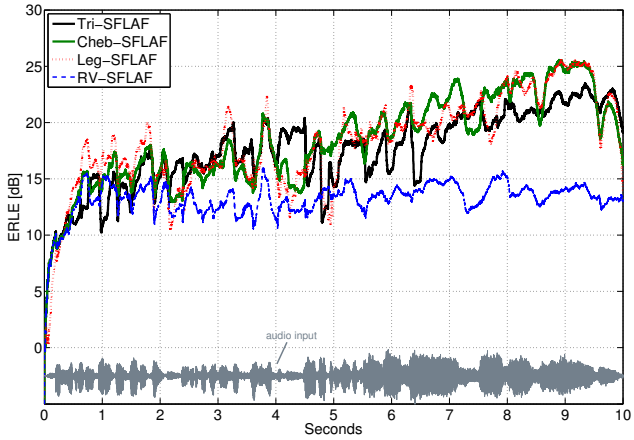

Fig. 8. Performance comparison in terms of ERLE between SFLAFs with different expansion types in case of audio input affected by an asymmetric sigmoidal nonlinearity.

TABLE V. PERFORMANCE COMPARISON IN TERMS OF OBJECTIVE MEASURES BETWEEN SFLAFS WITH DIFFERENT EXPANSION TYPES IN CASE OF AUDIO INPUT AFFECTED BY AN ASYMMETRIC SIGMOIDAL NONLINEARITY.

Measure	Tri-SFLAF	Che-SFLAF	Leg-SFLAF	RV-SFLAF
LLR	**0.077**	0.106	0.080	0.674
WSS	**4.76**	6.05	4.83	12.69
NCM	0.993	0.993	**0.998**	0.940

B. *Audio input*

We repeat the same experiments by applying the distortion to a different input signal, recorded from a radio station and composed of speech and music, thus showing different volume levels. Using the same configuration of the speech case, results in terms of the ERLE still show a superiority of the trigonometric SFLAF, which is more clear when the volume level of the input signal is higher, as depicted in Fig. 7. This result is confirmed also by quality and intelligibility indices in Table IV, where those indices that are exclusively intended for speech (i.e., PESQ and CSII) are no longer considered.

In the case of asymmetric sigmoidal nonlinearity (19), we find similar results as for speech input. In fact, as represented in Fig. 8, trigonometric, Chebyshev and Legendre SFLAFs alternate themselves as best model, although in Table V the quality indices (i.e., LLR and WSS) show best values for the trigonometric SFLAF while the best intelligibility (NCM measure) is provided by the Legendre SFLAF. Even in this case, although not reported, the Chebyshev SFLAF achieves best results when the input buffer length is very small.

VII. CONCLUSIONS

In this paper, a comparison has been made between functional link-based models involving different nonlinear expansions for audio and speech input signals. In particular, we assess the performance of the following expansion types: trigonometric series, Chebyshev polynomial, Legendre polynomial and random vector. Experiments are conducted using a purely speech signal and an audio signal composed of speech

and music. Two different nonlinearities are applied to the input signals. Overall, results have shown an advantage of the SFLAF using trigonometric series expansion, since it provides the best perceived quality of the signal, even when the error-based performance was the same as that obtained by the other expansions. Moreover, it involves the lowest computational cost. However, it is worth noting that, if we keep the input buffer length very small, good performance is achieved by the Chebyshev expansion, and this might be a significant result for those applications that require short filters.

ACKNOWLEDGMENT

The work of Danilo Comminiello was partly funded by bdSound.

REFERENCES

[1] C.-T. Tan, B. C. J. Moore, and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *J. Audio Engineering Society (AES)*, vol. 51, no. 11, pp. 1012–1231, Nov. 2003.

[2] A. Uncini, A. Nalin, and R. Parisi, "Acoustic echo cancellation in the presence of distorting loudspeakers," in *Proc. European Signal Processing Conference (EUSIPCO)*, vol. 1, Tolouse, France, 2002, pp. 535–538.

[3] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1502–1512, Jul. 2013.

[4] H. Schurer, C. H. Slump, and O. E. Herrmann, "Second order Volterra inverses for compensation of loudspeaker nonlinearity," in *Proc. IEEE ASSP Works. Appl. Signal Process. to Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 1995, pp. 74–78.

[5] A. Guerin, G. Faucon, and R. Le Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 672–683, Nov. 2003.

[6] A. Carini and G. L. Sicuranza, "Fourier nonlinear filters," *Signal Processing*, vol. 94, pp. 183–194, Jan. 2014.

[7] H. Dai and W.-P. Zhu, "Compensation of loudspeaker nonlinearity in acoustic echo cancellation using raised-cosine function," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 11, pp. 1190–1194, Nov. 2006.

[8] M. Scarpiniti, D. Comminiello, R. Parisi, and A. Uncini, "Nonlinear spline adaptive filtering," *Signal Processing*, vol. 93, no. 4, pp. 772–783, Apr. 2013.

[9] ——, "Hammerstein uniform cubic spline adaptive filters: Learning and convergence properties," *Signal Processing*, vol. 100, no. 7, pp. 112–123, Jul. 2014.

[10] G. L. Sicuranza and A. Carini, "A generalized FLANN filter for nonlinear active noise control," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2412–2417, Nov. 2011.

[11] J. C. Patra, R. N. Pal, B. N. Chatterji, and G. Panda, "Identification of nonlinear dynamic systems using functional link artificial neural networks," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 29, no. 2, pp. 254–262, Apr. 1999.

[12] D. Comminiello, L. A. Azpicueta-Ruiz, M. Scarpiniti, A. Uncini, and J. Arenas-García, "Functional link based architectures for nonlinear acoustic echo cancellation," in *Proc. IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, May 30 - Jun. 1 2011, pp. 180–184.

[13] J. C. Patra and A. C. Kot, "Nonlinear dynamic system identification using Chebyshev functional link artificial neural networks," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 32, no. 4, pp. 505–511, Aug. 2002.

[14] H. Zhao and J. Zhang, "Functional link neural network cascaded with Chebyshev orthogonal polynomial for nonlinear channel equalization," *Signal Processing*, vol. 88, no. 8, pp. 1946–1957, Aug. 2008.

[15] J. C. Patra, W. C. Chin, P. K. Meher, and G. Chakraborty, "Legendre-FLANN-based nonlinear channel equalization in wireless communication system," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, Singapore, Oct. 2008, pp. 1826–1831.

[16] A. Carini, S. Cecchi, L. Romoli, and G. L. Sicuranza, "Legendre nonlinear filters," *Signal Processing*, vol. 109, pp. 84–94, Apr. 2015.

[17] Y.-H. Pao, *Adaptive Pattern Recognition and Neural Networks*. Reading, MA: Addison-Wesley, 1989.

[18] Y.-H. Pao, G.-H. Park, and D. J. Sobajic, "Learning and generalization characteristics of the random vector functional-link net," *Neurocomputing*, vol. 6, no. 2, pp. 163–180, 1994.

[19] B. Igelnik and Y.-H. Pao, "Stochastic choice of basis functions in adaptive function approximation and the functional-link net," *IEEE Trans. Neural Networks*, vol. 6, no. 6, pp. 1320–1329, 1995.

[20] S. Scardapane, D. Wang, M. Panella, and A. Uncini, "Distributed learning for random vector functional-link networks," *Information Sciences*, vol. 301, pp. 271–284, Apr. 2015.

[21] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 42, no. 2, pp. 513–529, 2012.

[22] S. Scardapane, D. Comminiello, M. Scarpiniti, and A. Uncini, "Online sequential extreme learning machine with kernels," *IEEE Trans. Neural Networks and Learning Syst. (to appear)*, 2015.

[23] ——, "Music classification using extreme learning machines," in *8th Int. Symposium on Image and Signal Processing and Analysis (ISPA)*, Trieste, Italy, Sep. 2013, pp. 377–381.

[24] S. Scardapane, R. Fierimonte, D. Wang, M. Panella, and A. Uncini, "Distributed music classification using random vector functional-link nets," in *Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN)*, Killarney, Ireland, Jul. 2015.

[25] A. Uncini, *Fundamentals of Adaptive Signal Processing*, ser. Signal and Communication Technology. Cham, Switzerland: Springer International Publishing AG, 2015, ISBN 978-3-319-02806-4.

[26] B. Bank, "Computationally efficient nonlinear Chebyshev models using common-pole parallel filters with the application to loudspeaker modeling," in *130th Audio Engineering Society (AES) Convention*, London, UK, May 13-16 2011.

[27] J. M. Gil-Cacho, M. Signoretto, T. Van Waterschoot, M. Moonen, and S. H. Jensen, "Nonlinear acoustic echo cancellation based on a sliding-window leaky kernel affine projection algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1867–1878, Sep. 2013.

[28] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "Convergence properties of nonlinear functional link adaptive filters," *IET Electronics Letters*, vol. 49, no. 14, pp. 873–875, Jul. 2013.

[29] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Nonlinear acoustic echo cancellation based on sparse functional link representations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1172–1183, Jul. 2014.

[30] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T Recommend. P. 862 Std., 2000.

[31] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Salt Lake City, UT, May 2001, pp. 749–752.

[32] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[33] S. R. Quackenbush, S. R. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. NJ: Prentice Hall, 1988.

[34] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Paris, France, May 1982, pp. 1278–1281.

[35] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoustical Society of America*, vol. 117, no. 5, pp. 2224–2237, May 2004.

[36] I. Hollube and K. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, Sep. 1996.