Combined Nonlinear Echo Cancellation and Residual Echo Suppression

Andreas Schwarz, Christian Hofmann, Walter Kellermann

Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
Email: {schwarz,hofmann,wk}@lnt.de
Web: www.lms.lnt.de

Abstract

We describe a combined nonlinear acoustic echo cancellation and residual echo suppression system. The echo canceler uses parallel Hammerstein branches consisting of fixed nonlinear basis functions and linear adaptive filters. The residual echo suppressor uses an Artificial Neural Network for modeling of the residual echo spectrum from spectral features computed from the far-end signal. We show that modeling nonlinear effects both in the echo canceler and in the echo suppressor leads to an increased performance of the combined system.

1 Introduction

The conventional solution for echo cancellation in speech communication devices is a linear acoustic echo canceler (AEC), which models the acoustic path between loudspeaker output and microphone input with a linear filter, and subtracts the echo replica from the microphone signal [1]. For small devices producing high sound pressure levels, such as speakerphones or portable devices with voice control, this task is often complicated by nonlinear distortion and vibration effects which occur in the acoustic system and which cannot be modeled by linear echo cancelers [2]. This problem is even more relevant today with the increased use of mobile phones in speakerphone mode, due to the very small loudspeaker and enclosure dimensions, which lead to a high amount of nonlinear distortion.

Within the last decades, various approaches have been proposed for nonlinear acoustic system identification for echo cancellation. These range from the powerful, yet computationally expensive Volterra filters [3], for which more efficient approximations [4] and self-configuring structures [5, 6] have been developed, over time-variant selection and input-level-adaptive linear models [7] to low-complexity memoryless preprocessors [8–10]. Recently, also particle-filter algorithms have been successfully employed for the estimation of the parameters of a nonlinear dynamical system for acoustic echo cancellation [11]. On the one hand, all these echo cancelers subtract a phase-exact estimate of the acoustic echo from the microphone signal, which is why echo cancelers can perform very accurately, as far as the physical system can be approximated by the model. On the other hand, echo cancelers are also strictly limited to the deterministic mechanisms considered in their design phase and require precise adaptation to the actual acoustic environment. In practice, the effects caused by nonlinearities and vibrations cannot be modeled completely, such that they appear as random signal components [2] with characteristics depending on the input signal. For this reason, the AEC usually requires a residual echo suppressor (RES), realized as short-time spectral magnitude modification, e.g., using Wiener filtering or spectral subtraction [12]. Since, unlike the AEC, RES applies time-variant spectral weights in the microphone signal path, this will generally introduce near-end speech distortion, but allows a significantly higher degree of echo reduction than AEC alone, because coarse models in the short-time spectral magnitude domain can be used.

For cases where AEC filter length or convergence time are the limiting factors, the residual echo spectrum is still strongly correlated to the far-end signal, so that linear models for estimating the residual echo magnitude spectrum can be employed successfully. Linear models have also been applied for nonlinear echo paths, based on the observation that there is still some correlation between far-end signal and residual echo magnitude spectra [13, 14]. Also, models for harmonics in the time domain [15] or in the frequency domain [16] have been proposed. Recently, we proposed an echo suppressor using a spectral feature-based regression model, which models the residual echo as a function of low-dimensional features computed from the far-end magnitude spectrum [17].

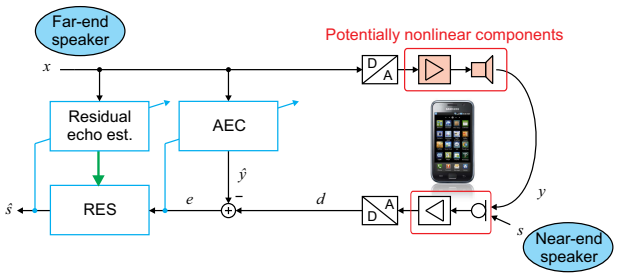

Figure 1: System structure for combined acoustic echo cancellation (AEC) and residual echo suppression (RES).

In this paper, we present a real-time-capable nonlinear echo reduction system consisting of a nonlinear AEC using a Hammerstein Group Model structure (similar as in [18–20]) with a robust frequency-domain adaptation algorithm in combination with spectral feature-based RES [17]. We first describe the overall structure of the system, introduce the non-linear AEC, and describe the implementation of the RES. Finally, we show results of an evaluation conducted with real smartphone recordings, considering echo return loss enhancement (ERLE) and signal distortion of the proposed system, as well as the modeling accuracy of the residual echo model.

2 System Description

Figure 1 shows the structure of a combined AEC and RES system. The microphone signal $d(n)$, where n is the discrete-time sample index, is composed of the desired near-end speech $s(n)$ and a linearly filtered and nonlinearly distorted version $y(n)$ of the far-end signal $x(n)$:

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

The AEC output $e(n)$ contains the near-end speech $s(n)$ and the residual echo $z(n)$ that remains after subtracting the echo estimate $\hat{y}(n)$:

$$e(n) = s(n) + y(n) - \widehat{y}(n) = s(n) + z(n). \tag{2}$$

For the RES, the AEC output signal $e(n)$ and the far-end signal $x(n)$ are decomposed using a uniform analysis filter bank, yielding the frequency-subband signals $E(\nu,k)$ and $X(\nu,k)$, respectively, with the frequency index ν and the frame-time index k. In the following, we will omit the time index k whenever possible. The filter bank is characterized by an FIR prototype filter of length L, DFT size K, and frame shift N_s. The filter bank output vector capturing all subband signal samples at

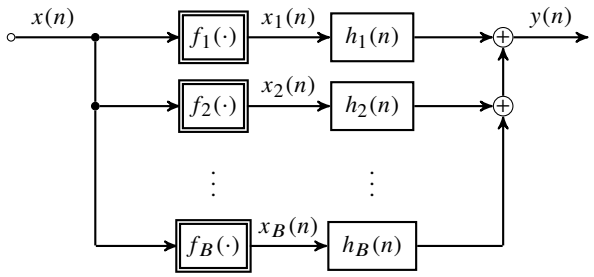

Figure 2: Block diagram of a Hammerstein group model with B branches, each of which is composed of a Hammerstein model.

a given time k has $N_B = K/2+1$ unique complex coefficients and is denoted as *spectrum* in the following. The magnitude spectra of the AEC output and the reference signal are defined by $M_E(\nu,k) = \mathcal{E}\{|E(\nu,k)|\}$ and $M_X(\nu,k) = \mathcal{E}\{|X(\nu,k)|\}$, respectively, where the expectation $\mathcal{E}$ is realized in practice by recursive temporal smoothing with a forgetting factor λ close to 1. The RES applies a frequency-dependent gain $G(\nu,k)$ to the AEC output signal to obtain an estimate for the near-end signal $\widehat{S}(\nu,k) = G(\nu,k)E(\nu,k)$, which is re-synthesized into a time-domain signal $\hat{s}(n) = s_{\text{out}}(n) + z_{\text{out}}(n)$ consisting of the potentially distorted and attenuated near-end speech component $s_{\text{out}}(n)$ and the remaining residual echo component $z_{\text{out}}(n)$.

2.1 Nonlinear Acoustic Echo Cancellation

A very simple echo-path model is an adaptive causal linear finite-impulse-response (FIR) system. Such a system is completely described by

$$\hat{y}(n) = \sum_{\kappa=0}^{L_h-1} \hat{h}(\kappa)x(n-\kappa) = x(n) * \hat{h}(n), \tag{3}$$

where n is the discrete-time sample index, $x(n)$ is the input signal, $\hat{h}(n)$ is the system's estimated impulse response of length L_h, and where $*$ denotes linear convolution. For practical applications, the filter coefficients of such models are typically adapted by LMS-type algorithms, such as the normalized least-mean-square (NLMS) algorithm, or by affine projection or recursive least-squares (RLS) algorithms (see [21] for an extensive review of such algorithms).

A simple nonlinear echo-path model is a Hammerstein model, consisting of a memoryless nonlinearity and a subsequent linear system. Such a structure is justified as an approximation of the cascade of nonlinearly distorting playback equipment followed by the linear propagation of the radiated sound waves through the room to a microphone. Hammerstein group models (HGMs) are comprised of a group of B parallel Hammerstein models, denoted as B branches [20]. The block diagram corresponding to this structure is depicted in Fig. 2, for which the input-output relation can be written as

$$y(n) = \sum_{b=1}^{B} x_b(n) * h_b(n), \tag{4}$$

where b is the branch index and $x_b(n) = f_b\{x(n)\}$ is the b^{th} branch signal, $h_b(n)$ is called the linear kernel, and $f_b(\cdot)$ the nonlinear base function of branch b.

Note that traditionally employed HGMs are the so-called power filters [18]. More recently, Fourier-base HGMs have been proposed in [19] and HGMs with Legendre polynomials have been employed in [20]. Furthermore, note that power filters are the special case of Volterra filters where only the main diagonal of each Volterra kernel is non-zero, and that Legendre-base HGMs can be equivalently expressed as power filters of appropriate orders and vice versa.

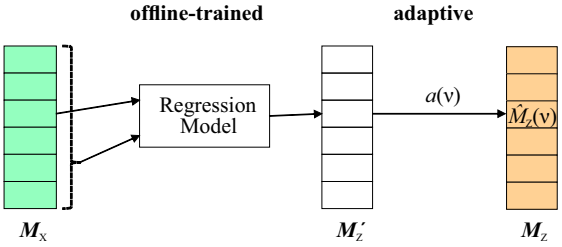

Figure 3: Residual echo suppressor combining an offline-trained regression model with adaptive scalar weights.

As the output of an HGM linearly depends on its kernel coefficients $h_b(n)$ and the branch signals $x_b(n)$, classical multi-channel algorithms for linear systems can be employed to adapt the set of $\hat{h}_b(n)$ of an adaptive HGM for a predefined set of base functions $f_1(\cdot),\ldots,f_B(\cdot)$ to model a physical system, employing the physical system's input and output signals (microphone signals). In particular, we use the robust frequency-domain adaptive filter proposed in [22] to adapt the branch kernels. The algorithm employs a Newton-Raphson iteration for the filter update to achieve fast convergence and achieves robustness against disturbances by employing a Huber function instead of a squared-error criterion. A correlation-based double-talk detector (DTD) employing a quickly-adapting linear shadow filter is used [22, 23]. In the context of nonlinear system identification, branch-specific step sizes can be chosen for the adaptation of the HGM to emphasize the contribution of the linear model and prevent over-adaptation of the nonlinear subsystem to the input signal.

2.2 Residual Echo Suppression

The task of the residual echo suppression is the computation of the gain G, based on the estimated magnitude spectrum $\widehat{M}_Z$ of the residual echo $z(n)$, and the AEC output signal magnitude M_E. We employ the Wiener filter rule

$$G(\nu) = \max\left(G_{\min}, 1-\mu\frac{\widehat{M}_Z^2(\nu)}{M_E^2(\nu)}\right), \tag{5}$$

with the overestimation factor μ and the minimum gain $G_{\min}$. It is clear that the remaining problem of residual echo suppression is the estimation of the residual echo magnitude spectrum $\widehat{M}_Z(\nu)$. To this end, the residual echo magnitude spectrum can be modeled as a function of the magnitude spectrum of a reference signal, here, the far-end signal x ($\widehat{y}$ has also been used [13]).

The method that we employ in this paper has been proposed in [17] and will be briefly reviewed in the following. The structure is illustrated in Fig. 3. The first stage is a regression model, which, in each subband, uses the magnitude of the same subband of the reference signal, as well as one or more spectral features which are computed from the reference signal, to obtain an initial estimate $\widehat{M}_Z'(\nu,k)$, i.e.,

$$i_1(\nu,k) = M_X(\nu,k), \tag{6}$$

$$i_2(\nu,k) = \frac{1}{\nu/2}\sum_{m=1}^{\nu/2} M_X(m,k), \tag{7}$$

$$\widehat{M}_Z'(\nu,k) = R_\nu(i_1(\nu,k), i_2(\nu,k)). \tag{8}$$

Here, i_2 is a feature computed by averaging over the magnitudes of all reference subbands up to half of the subband ν for which M_Z is to be estimated, so that all subharmonics are captured. The regression model is implemented as an artificial neural network and trained offline on representative residual echo signals

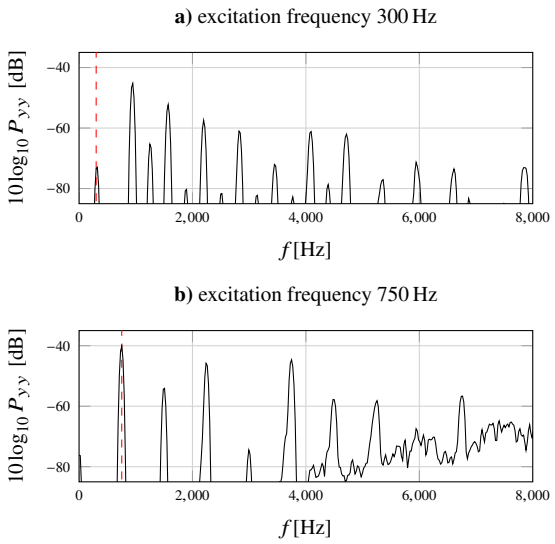

Figure 4: Echo spectrum for loudspeaker excitation with a single sine wave; maximum amplitude, same playback gain as used for speech signals.

recorded with the device. The second stage consists of multiplying an adaptive scalar parameter $a(\nu,k)$ in each subband to obtain the final estimate

$$\widehat{M}_Z(\nu,k) = a(\nu,k)\widehat{M}'_Z(\nu,k). \tag{9}$$

The parameter $a(\nu,k)$ is adapted using the update equation

$$a(\nu,k) = a(\nu,k-1) + \mu_a\Big(M_E(\nu,k) - a(\nu,k)\widehat{M}'_Z(\nu,k)\Big) \tag{10}$$

whenever the DTD of the AEC indicates that near-end speech and noise are negligible. In this way, the estimate from the fixed regression model is continuously adapted to the current acoustic conditions.

3 Evaluation

For a realistic evaluation of the proposed system, we use signals recorded with a commercial smartphone with a 4.7 inch screen diagonal. The device has a microphone on the top edge and a single speaker port on the back near the bottom.

For training the RES models, we use an echo signal of 30 s duration containing male and female speech recorded with the device in an anechoic environment. The echo signals for evaluation consist of different male and female speech signals, recorded in a reverberant environment with $T_{60} \approx 0.4$ s, for 5 different recording conditions (device placed in different orientations and on different surfaces). The playback gain of the device was set to yield a sound pressure level of about 70 dB(A) at 1 m distance, which causes strongly audible nonlinear distortions. For the evaluation of double talk performance, near-end speech signals recorded with the phone are added to the recorded echo, with a near-end to echo ratio of about -6 dB.

Fig. 4 shows two examples for the echo spectrum resulting from excitation of the loudspeaker with a single sine wave. For 300 Hz, odd-order harmonics are dominant, exceeding even the linear echo component; these can be effectively reduced by the nonlinear echo canceler. For 750 Hz, noise-like effects occur, which can not be modeled by the echo canceler, but require echo suppression.

The nonlinear AEC in the experiments employs an adaptive HGM with Legendre polynomials of orders 1 (linear), 3 and 5 as base functions. Adaptation is performed with the aforementioned robust frequency-domain algorithm, with a stepsize that is lower by a factor of two for the nonlinear branches compared to the linear branch. All filters have the length $L_h = 512$. For the proposed residual echo suppressor, we use a feed-forward artificial neural network with 2 hidden layer nodes, which is trained using the Levenberg-Marquardt algorithm [24] with a mean square error cost function, followed by the adaptive stage controlled by the AEC DTD. As a baseline for comparison, we use a RES with the same structure and adaptation procedure, but employing, instead of the neural network regression model, a model consisting of fixed scalar weights as the first stage, where the weights are optimized for minimum MSE on the training signals.

The computational complexity of the proposed system is only moderately increased over the baseline system. The NL-AEC complexity is about twice as high as for linear AEC, while the NL-RES requires an additional 6 multiplications and 2 evaluations of the log-sigmoid function (which can be efficiently implemented using a lookup table) per frame and subband.

For the residual echo suppression, the parameters of the analysis-synthesis filter bank are set to $L = 512$, $K = 256$ and $N_s = 64$, i.e., we have $N_B = 129$ non-redundant subbands, and the prototype filter coefficients are computed according to [25]. The recursive smoothing parameter for the magnitude spectrum estimation is set to $\lambda = 0.92$.

To evaluate the accuracy of the residual echo modeling we compute the relative MSE between the estimated and measured residual echo spectrum:

$$\mathrm{MSE}_{\mathrm{rel}} = \frac{\sum_{\nu,k}(M_Z(\nu,k) - \hat{M}_Z(\nu,k))^2}{\sum_{\nu,k} M_Z^2(\nu,k)}. \tag{11}$$

For the echo reduction performance, we evaluate the ERLE of the AEC and the combined AEC and RES:

$$\mathrm{ERLE}_{\mathrm{AEC}} = 10\log_{10}\frac{\mathcal{E}\{y^2\}}{\mathcal{E}\{z^2\}}, \tag{12}$$

$$\mathrm{ERLE}_{\mathrm{AEC,RES}} = 10\log_{10}\frac{\mathcal{E}\{y^2\}}{\mathcal{E}\{z_{\mathrm{out}}^2\}}, \tag{13}$$

where the expectation operator $\mathcal{E}$ is approximated by temporal averaging over the echo only (single-talk, ST) or double-talk (DT) periods of the evaluation signal, skipping the initial convergence phase (7.5 s). The undesired distortion to the near-end signal caused by the RES in the double-talk case is quantified by the near-end signal attenuation (NEA) and the segmental signal to distortion ratio (SSDR):

$$\mathrm{NEA} = 10\log_{10}\frac{\mathcal{E}\{s^2\}}{\mathcal{E}\{s_{\mathrm{out}}^2\}}, \mathrm{SSDR} = \mathrm{SSNR}(s, s - s_{\mathrm{out}}), \tag{14}$$

where SSNR is the segmental SNR averaged over segments of 256 samples, with the segment SNR limited to the range -10 dB…35 dB [26].

In Fig. 5, we illustrate the evaluation signal for one recording condition, showing the residual echo (red) and near-end (black) signal components for the microphone signal, the linear AEC output, the NL-AEC output, and the proposed combination of nonlinear AEC and RES. The corresponding audio files are available online [27]. Table 1 shows the results of the performance measures, averaged over all 5 recording conditions. The nonlinear AEC alone significantly improves the ERLE values; in combination with the proposed RES, ERLE is further improved. Additionally, near-end signal attenuation caused by the RES is lowered if the nonlinear AEC is used. Furthermore, it is worth noting that the relative modeling error (relative MSE) for the residual echo after nonlinear AEC is lower than after linear AEC, which confirms that the proposed RES model is particularly effective in modeling effects that cannot be modeled by nonlinear AEC.

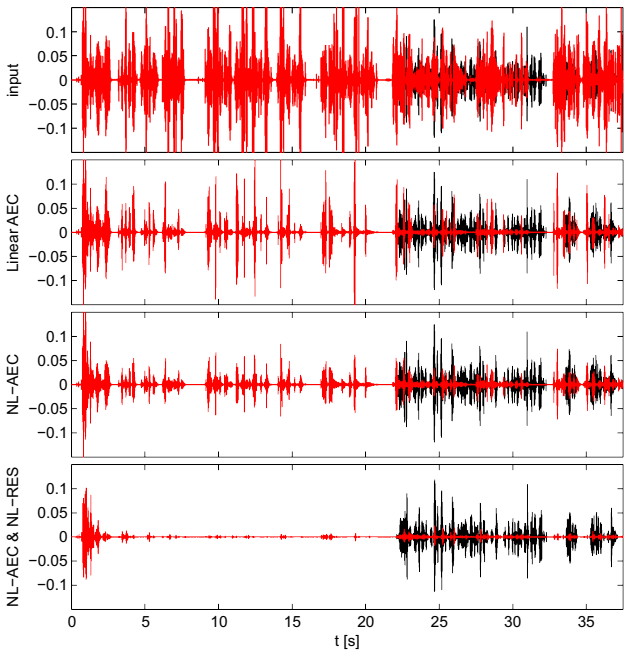

Figure 5: Example for echo (red) and near-end (black) signal components at input (microphone) and after AEC and RES.

AEC	RES	MSE_{rel}	$ERLE_{ST}$ [dB]	$ERLE_{DT}$ [dB]	NEA [dB]	SSDR [dB]
	-	-	10.8	10.0	-	-
Linear	baseline	0.50	15.9	14.8	0.30	14.6
	NL-RES	0.33	27.1	21.3	0.59	13.1
	-	-	14.7	12.6	-	-
NL-AEC	baseline	0.48	20.3	17.1	0.24	16.1
	NL-RES	0.31	30.0	21.5	0.51	13.4

Table 1: AEC and RES performance measures.

4 Conclusions

We have shown results for a combined AEC and RES system, where nonlinear effects are considered both in the HGM-AEC and in the spectral feature-based RES stage. We found that the employed residual echo estimator benefits from modeling of harmonics in the NL-AEC, as indicated by the lower relative modeling error for the residual. Due to its effectiveness and low complexity, the proposed combination is a promising approach for implementation in mobile devices.

References

[1] C. Breining, P. Dreiseitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. an application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16(4), pp. 42–69, July 1999.

[2] A. Birkett and R. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. WASPAA*, 1995.

[3] A. Stenger and R. Rabenstein, "Adaptive Volterra filters for nonlinear acoustic echo cancellation," in *Proc. NSIP*, 1999.

[4] W. A. Frank, "An efficient approximation to the quadratic Volterra filter and its application in real-time loudspeaker linearization," *Signal Processing*, vol. 45, no. 1, pp. 97 – 113, 1995.

[5] M. Zeller, L. Azpicueta-Ruiz, J. Arenas-Garcia, and W. Kellermann, "Adaptive Volterra filters with evolutionary quadratic kernels using a combination scheme for memory control," *IEEE Trans. Signal Processing*, vol. 59, pp. 1449–1464, April 2011.

[6] M. Zeller and W. Kellermann, "Evolutionary adaptive filtering based on competing filter structures," in *Proc. EUSIPCO*, 2011.

[7] S. Saito, A. Nakagawa, and Y. Haneda, "Dynamic impulse response model for nonlinear acoustic system and its application to acoustic echo canceller," in *Proc. WASPAA*, 2009.

[8] A. Stenger and R. Rabenstein, "An acoustic echo canceller with compensation of nonlinearities," in *Proc. EUSIPCO*, 1998.

[9] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, no. 9, pp. 1747–1760, 2000.

[10] S. Shimauchi and Y. Haneda, "Nonlinear acoustic echo cancellation based on piecewise linear approximation with amplitude threshold decomposition," in *Proc. IWAENC*, 2012.

[11] C. Huemmer, C. Hofmann, R. Maas, A. Schwarz, and W. Kellermann, "The elitist particle filter based on evolutionary strategies as novel approach for nonlinear acoustic echo cancellation," in *Proc. ICASSP*, 2014.

[12] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64(1), pp. 21–32, Jan. 1998.

[13] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," in *Proc. ICASSP*, 2006.

[14] O. Hoshuyama, "An update algorithm for frequency-domain correlation model in a nonlinear echo suppressor," in *Proc. IWAENC*, 2012.

[15] F. Kuech and W. Kellermann, "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," in *Proc. ICASSP*, 2007.

[16] D. Bendersky, J. Stokes, and H. Malvar, "Nonlinear residual acoustic echo suppression for high levels of harmonic distortion," in *Proc. ICASSP*, 2008.

[17] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *Proc. WASPAA*, 2013.

[18] F. Kuech, A. Mitnacht, and W. Kellermann, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *Proc. ICASSP*, 2005.

[19] S. Malik and G. Enzner, "Fourier expansion of Hammerstein models for nonlinear acoustic system identification," in *Proc. ICASSP*, 2011.

[20] C. Hofmann, C. Huemmer, and W. Kellermann, "Significance-aware Hammerstein group models for nonlinear acoustic echo cancellation," in *Proc. ICASSP*, 2014.

[21] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River (NJ), USA: Prentice Hall, 4th ed., 2002.

[22] H. Buchner, J. Benesty, T. Gaensler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Trans. ASLP*, vol. 14(5), pp. 1633–1644, Sept. 2006.

[23] T. Gänsler, S. Gay, M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 656–663, Nov 2000.

[24] M. Hagan and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Networks*, vol. 5(6), pp. 989–993, Nov. 1994.

[25] M. Harteneck, S. Weiss, and R. Stewart, "Design of near perfect reconstruction oversampled filter banks for subband adaptive filters," *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46(8), pp. 1081–1085, Aug. 1999.

[26] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP*, 1998.

[27] `http://www.lms.lnt.de/files/publications/itgspeech2014-nonlinear.zip`.