2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 20-23, 2013, New Paltz, NY

SPECTRAL FEATURE-BASED NONLINEAR RESIDUAL ECHO SUPPRESSION

Andreas Schwarz, Christian Hofmann, Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{schwarz, hofmann, wk}@LNT.de

ABSTRACT

We propose a method for nonlinear residual echo suppression that consists of extracting spectral features from the far-end signal, and using an artificial neural network to model the residual echo magnitude spectrum from these features. We compare the modeling accuracy achieved by realizations with different features and network topologies, evaluating the mean squared error of the estimated residual echo magnitude spectrum. We also present a low complexity real-time implementation combining an offline-trained network with online adaptation, and investigate its performance in terms of echo suppression and speech distortion for real mobile phone recordings.

***Index Terms*—** Nonlinear acoustic echo suppression, residual echo suppression, AES, RES

1. INTRODUCTION

Echo cancellation is a well-known problem in acoustic signal processing in telecommunications. The conventional solution is a linear acoustic echo canceler (AEC), which models the loudspeaker-enclosure-microphone signal path with a linear filter, and subtracts the echo replica from the microphone signal [1]. For speakerphones, the problem 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 severe with today's mobile phones in speakerphone mode, due to the very small loudspeaker and enclosure dimensions, which lead to a high amount of nonlinear distortion. Various approaches have been proposed for nonlinear acoustic system identification for echo cancellation, ranging from universal but computationally costly Volterra filters [3] over many alternatives with intermediate complexity [4] to a simple memoryless polynomial preprocessor followed by a linear filter [5]. However, common to all these approaches is that they require significantly more computational effort than linear echo cancelers, and that they can only model deterministic effects of the acoustic system, not the noise-like artifacts which also occur in practice [2].

Due to its limited modeling capability, the AEC is typically augmented with a residual echo suppressor (RES), realized as a frequency-domain Wiener filter or spectral subtraction [6]. This approach will generally introduce near-end speech distortion, but allows a significantly higher degree of echo reduction than AEC alone. 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, therefore linear models for estimating the residual echo magnitude spectrum can be used successfully. Linear models have also been applied for nonlinear echo paths, based on the observation that some correlation between far-end signal and residual echo magnitude spectra exists [7, 8]. Also, models for harmonics in the time domain [9] or in the frequency domain [10] have been proposed.

In this paper, we propose a spectral feature-based model for the estimation of the residual echo magnitude spectrum. Due to the complexity of the physical processes leading to distortion artifacts, we do not attempt to model these processes directly. Instead, we extract features from the far-end signal spectrum, and train a multiple-input regression model, realized as an artificial neural network, for the estimation of the residual echo magnitude spectrum. For online adaptation, we combine the offline-trained regression model with online-adapted linear weights.

This paper is structured as follows: we first describe the structure of a typical combined AEC and RES system. We then review existing models for estimating residual echo spectra, and introduce our spectral feature-based modeling approach. We show how the approach can be used in a low-complexity real-time implementation. Finally, we evaluate the performance of the proposed model, and show results of the real-time implementation.

2. RESIDUAL ECHO SUPPRESSION

Fig. 1 shows the structure of a combined AEC and RES system. The microphone signal $d(t)$ comprises the desired near-end signal $s(t)$ and a linearly filtered and nonlinearly distorted version $y(t)$ of the far-end signal $x(t)$:

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

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

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

For the RES, the AEC output signal $e(t)$ and the far-end signal $x(t)$ are decomposed using a uniform analysis filter bank, yielding the frequency-subband signals $E(\nu, k)$ and $X(\nu, k)$, with the frequency index ν and the 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 with length L, DFT size K, and frame shift N_s. The filter bank output vector capturing all subband signal samples at a given time k has $N_\mathrm{B} = K/2+1$ unique complex coefficients and is denoted as spectrum in the following. We estimate the

The authors thank Samsung Electronics Co. for supporting this work.

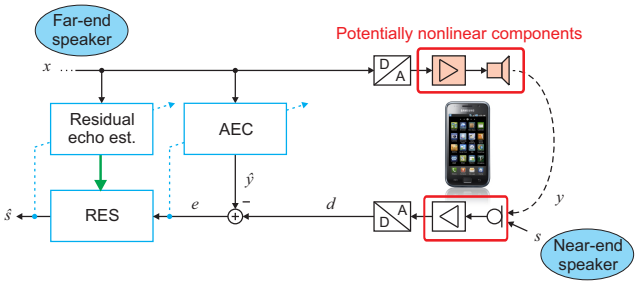

Figure 1: Signal model for acoustic echo cancellation (AEC) and residual echo suppression (RES)

magnitude spectra of the AEC output $M_E(\nu,k)$ and the reference signal $M_X(\nu,k)$ by recursive temporal smoothing with a forgetting factor λ close to 1. The RES applies a frequency-dependent gain to the AEC output signal to obtain an estimate for the near-end signal $\widehat{S}(\nu) = G(\nu)E(\nu)$. The MMSE-optimal suppression gain G is computed based on $\widehat{M}_Z$, representing the estimated magnitude spectrum of the residual echo $z(t)$, and the AEC output signal magnitude M_E, using the Wiener filter rule

$$G(\nu) = \max\left(G_{\min}, 1 - \mu\frac{\widehat{M}_Z^2(\nu)}{M_E^2(\nu)}\right), \tag{3}$$

with the overestimation factor μ and the minimum gain $G_{\min}$. The estimate $\widehat{S}$ is then transformed back into a time-domain fullband signal using a synthesis filter bank, yielding the output signal $\widehat{s}$, which consists of the processed and potentially distorted near-end signal s_{out} and a remaining residual echo component z_{out}:

$$\widehat{s}(t) = s_{\text{out}}(t) + z_{\text{out}}(t). \tag{4}$$

3. NONLINEAR RESIDUAL ECHO MAGNITUDE SPECTRUM ESTIMATION

The core problem of residual echo suppression is the estimation of the residual echo magnitude spectrum $\widehat{M}_Z(\nu)$. In the following, we only consider frequency-domain estimation methods, where the magnitude spectrum of the residual echo is modeled as a function of the magnitude spectrum of a reference signal. We use the far-end signal x as the reference signal, while, e.g., in [7], the echo estimate of the AEC $\widehat{y}$ is used.

3.1. Linear Model

First, we consider a simple model where the residual echo magnitude in each subband is modeled from the corresponding reference signal subband magnitude multiplied by a frequency-dependent scalar parameter $a(\nu)$. We refer to this as the "linear model":

$$\widehat{M}_{Z,\text{Sc}}(\nu) = a(\nu)M_X(\nu). \tag{5}$$

The most attractive characteristic of the linear model is the low number of only N_{B} parameters, which can be estimated using linear regression. Although the model implicitly assumes a linear echo path [6], where each subband is only dependent on the same subband of the reference signal, this approach has also explicitly been proposed and is widely used for nonlinear echo suppression in mobile phones [7, 8]. However, even given perfect parameter estimation, this model has the fundamental limitation that it does not take into account coupling between subbands, as it is expected for nonlinear echo paths.

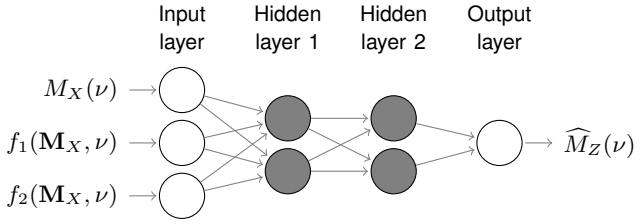

Figure 2: Example of network structure proposed for residual echo magnitude spectrum estimation

3.2. Full Coupling Model

As a generalization of the linear model, we consider an estimator that models each output subband as a linear combination of all input subbands, i.e., the entire vector $\mathbf{M_X} = [M_X(1)\dots M_X(N_{\text{B}})]^T$:

$$\widehat{M}_{Z,\text{FC}}(\nu) = \mathbf{a}(\nu)^{\text{T}}\mathbf{M_X}. \tag{6}$$

We refer to this as the "full coupling model". We note that in [11], a weighted combination of several samples within a given subband signal (but not across subbands, as we propose here for nonlinear echo paths) was used for suppression of linear echos. As for the linear model above, the optimization of the parameter vectors $\mathbf{a}(\nu)$ is a linear regression problem. However, with N_{B}^2, the number of total parameters is squared compared to the linear model.

Intermediate models between the linear and the full coupling model are conceivable. In [10], only those subbands that can generate harmonics in the current subband are considered, corresponding to sparse coupling vectors $\mathbf{a}(\nu)$ in our model. However, this explicit model for harmonics does not account for the complex nonlinear effects which occur in practice [2].

3.3. Proposed Spectral Feature-Based Model

We propose a multiple-input nonlinear regression model for the relationship between the far-end signal magnitude spectrum and the residual echo magnitude spectrum, using spectral features extracted from the far-end signal magnitude spectrum, and an artificial neural network for the estimation of the residual echo magnitude spectrum from these features.

Fig. 2 shows the topology of the feedforward network that we propose for the estimation of the residual echo magnitude in each subband, consisting of an input layer, an arbitrary number of hidden layers, and an output layer. Each hidden layer node represents a log-sigmoid function which is applied to the weighted sum of the inputs and an additional bias value. The output layer node represents a weighted sum of the hidden layer outputs and an additional bias value. For the first input of the network, we use the magnitude of the reference signal in the current subband $M_X(\nu)$, which we know is strongly correlated to the output variable [7]. The other inputs are derived from $\mathbf{M}_X$ using any number of feature-generating functions, in this example, $f_1(\mathbf{M}_X)$, $f_2(\mathbf{M}_X)$. This model can be seen as a generalization of the simple linear relationship between the reference signal magnitude and the residual echo magnitude, and also of the sparse and full coupling models, which can be recreated by using the magnitudes of (some or all) other subbands as input features, and no hidden layer nodes, resulting in a purely linear combination of the inputs.

The purpose of the feature extraction functions is to provide the network with information from other subbands that affect the residual echo magnitude in the current subband. An example of a feature extraction function is the average over all subbands up to

half of the current subband ν, motivated by the observation that nonlinear components in subband ν are likely to result from input components in subbands $\frac{\nu}{2}$ and less, e.g., if they represent higher-order harmonics.

$$f_1(\mathbf{M}_X(k),\nu) = \frac{1}{\nu/2}\sum_{m=1}^{\nu/2} M_X(m,k) \quad (7)$$

An even simpler feature is the average over all subband magnitudes:

$$f_2(\mathbf{M}_X(k),\nu) = f_2(\mathbf{M}_X(k)) = \frac{1}{N_\mathrm{B}}\sum_{m=1}^{N_\mathrm{B}} M_X(m,k) \quad (8)$$

Other features created by varying the summation range and according normalization in (7) from $\frac{\nu}{2}$ to ν, $\nu+1$ and $\nu-1$ were evaluated, but did not lead to significantly different results.

4. REAL-TIME IMPLEMENTATION

Compared to the AEC or the filterbank implementation, the evaluation of the feedforward network for a given input is an operation of negligible complexity, requiring, e.g., for a network with two inputs and two hidden nodes, eight multiplications and two evaluations of the log-sigmoid function (which can be implemented efficiently as a look-up table) per subband. However, the training of a feedforward network is in general a non-convex optimization problem, therefore online adaptation of the network parameters is not feasible. For an efficient real-time implementation, we therefore propose to use an offline-trained network to obtain an initial residual echo magnitude estimate $\widehat{M}_z'(\nu,k)$, which is then weighted by a scalar factor $a(\nu,k)$ which is adapted online:

$$\widehat{M}_{Z,\mathrm{adapt}}(\nu,k) = a(\nu,k)\widehat{M}_Z'(\nu,k), \quad (9)$$

resulting in a combination of the linear model and the spectral feature-based model. The assumption behind this combination is that the nonlinear characteristics, i.e., the ratio between the linear and the nonlinear components in a subband, are not strongly dependent on the acoustic environment, and can therefore be modeled by the offline-trained network, while the effect of the time-variant acoustic environment is modeled by the adaptive weighting. The adaptive factor $a(\nu,k)$ should adapt to the ratio between the true residual echo magnitude and the echo magnitude $\widehat{M}_Z'^2(\nu,k)$ estimated by the offline-trained network. This is achieved by adapting the weights in echo-only periods (as indicated by the DTD of the AEC) using the LMS update

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

where μ_a is a step size parameter. As an alternative, the adaptation procedure proposed in [8] could be applied.

5. EVALUATION

For our evaluation of the modeling accuracy, we use echo signals of male and female speech, recorded with a current-generation smartphone in an anechoic environment (i.e., the echo consists mostly of the direct path between loudspeaker and microphone) at a sampling rate of 16 kHz, with a 120 Hz high-pass filter. First, we apply a linear, frequency-domain AEC [12] which is continuously adapted on the signals, with a filter length of 512 taps. The AEC alone yields an ERLE (echo return loss enhancement) of approx. 7 dB, limited by the high level of nonlinear distortion. The parameters of the

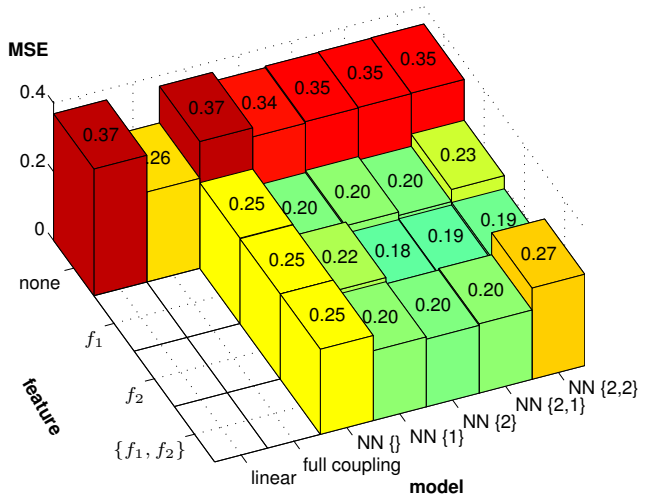

Figure 3: Modeling performance of linear, full coupling, and spectral feature-based models

analysis-synthesis filter bank are set to $L = 512$, $K = 128$ and $N_\mathrm{s} = 32$, i.e., we have $N_\mathrm{B} = 65$ unique subbands. The smoothing parameter is set to $\lambda = 0.95$. MMSE training of the RES models is performed on a 60 s period of the signals. The network parameters (weights and biases) of the proposed model are trained using the Levenberg-Marquardt algorithm [13], with the mean square error as objective function. We then evaluate the MSE between the estimated and measured residual echo spectrum by averaging over a 10 s period (not overlapping with the training period):

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

Fig. 3 shows the MSE obtained with the linear model, the full coupling model, and the spectral feature-based model in various configurations, where one axis indicates the features that are used as input to the network in addition to the magnitude of the current subband (none, f_1, f_2, $\{f_1, f_2\}$), and the other axis indicates the network structure that was used. The evaluated networks are a linear input combination with no biases (NN lin {}) and different nonlinear networks, from one hidden layer containing one or two log-sigmoid nodes (NN {1}/{2}) to two layers containing two nodes each (NN {2,2}). We can see that even the linear feature combination, which requires just one additional parameter per subband, significantly outperforms the linear model. Models using one or two hidden nodes show further increased performance. More hidden nodes or multiple hidden layers did not yield a consistent improvement, and often a reduction of modeling accuracy, indicating overfitting of the model on the given training data. Although the full coupling model shows almost the same performance as the linear feature combination here, we found that the generality of this model is worse, i.e., it performs worse when evaluated on signals recorded with different speakers or in a different environment than used for training, due to the large number of parameters N_B per subband.

For the real-time implementation, we evaluate the ERLE of the AEC and the combined AEC and RES:

$$\mathrm{ERLE}_\mathrm{AEC} = 10\log_{10}\frac{E\{y^2\}}{E\{z^2\}},\ \mathrm{ERLE}_\mathrm{AEC,RES} = 10\log_{10}\frac{E\{y^2\}}{E\{z_\mathrm{out}^2\}}, \quad (12)$$

where the expectation is realized by averaging. For the evaluation of the effect on the near-end signal, we compute the level of unde-

sired near-end signal attenuation (NEA) and the segmental signal to distortion ratio (SSDR):

$$\text{NEA} = 10\log_{10}\frac{E\{s^2\}}{E\{s_{\text{out}}^2\}},\ \text{SSDR} = \text{SSNR}(s, s - s_{\text{out}}), \quad (13)$$

where SSNR is the well-known segmental SNR, averaged over segments of 256 samples, where the segment SNR is limited to the range -10 dB. . . 35 dB [14]. The speech signals for the evaluation were recorded in a reverberant environment with $T_{60} \approx 0.3$ s. The playback volume of the phone is set to yield a sound pressure level of approx. 70 dB(A) in 1 m distance, which causes strongly audible nonlinear distortions. For the evaluation of double talk performance, we add near-end speech to the recorded echo, with a near-end to echo ratio of approx. 0 dB, and perform AEC and RES adaptation on the combined signal. The offline estimator for the proposed RES is a network with one log-sigmoid hidden node, chosen because the advantage of the two-node network in the MSE evaluation was not significant, and because a less complex model is usually advantageous with respect to generality. The feature f_1 (7) is used as the second input to the network. The network parameters are trained on the anechoic recordings described above. We compare the proposed real-time RES with a baseline system using the linear model and the same adaptation procedure. We set $\mu = 5$ and $G_{\text{min}} = 0$ for both systems. Compared to the baseline RES, the proposed RES requires a total of $3N_{\text{B}}$ additional multiplications and N_{B} log-sigmoid table lookups per analysis frame.

Fig. 4 shows the echo (red) and near-end (black) components within the microphone signal, the AEC output signal, the output of the baseline RES and of the proposed RES. Table 1 shows the overall ERLE during single talk (5-15 s) and double talk (15-37 s), as well as NEA and SSDR during double talk. Evidently, the proposed RES increases echo suppression without significantly increasing near-end distortion or attenuation. Increasing the suppression factor of the baseline RES to $\mu = 7.5$ leads to higher near-end attenuation and distortion, but still lower ERLE compared to our proposed RES, which confirms the advantage of the spectral feature-based model. Note that, in practice, additional frequency-independent suppression is commonly applied to achieve stronger echo reduction in single talk periods; this is not included in this evaluation for clarity of presentation.

	single talk	double talk		
	ERLE	ERLE	NEA	SSDR
AEC only	10.7 dB	10.6 dB	-	-
+basel. RES, $\mu = 5.0$	15.5 dB	16.1 dB	0.38 dB	13.4 dB
+basel. RES, $\mu = 7.5$	15.7 dB	17.0 dB	0.46 dB	12.0 dB
+prop. RES, $\mu = 5.0$	24.0 dB	19.2 dB	0.45 dB	13.0 dB

Table 1: RES performance for speech recorded with a mobile phone in a real environment

6. CONCLUSION

We have presented a spectral feature-based model for residual echo magnitude spectrum estimation, which shows increased modeling performance over the widely used linear model. We have furthermore presented a low-complexity real-time implementation of a system combining the offline-trained feature-based model with an online-adapted linear model, and demonstrated that it leads to increased echo suppression performance for a realistic mobile phone scenario with a high amount of nonlinear distortion.

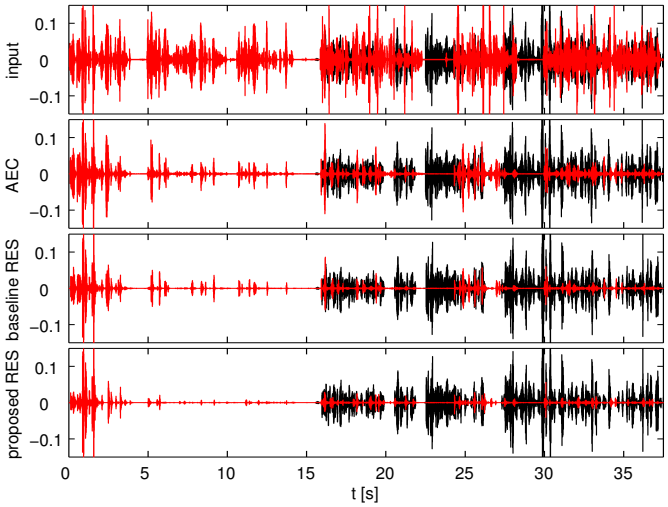

Figure 4: Residual echo (red) and near-end signal (black) components for baseline and proposed RES

7. 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 Magazine*, vol. 16(4), pp. 42–69, Jul. 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] S. Malik and G. Enzner, "Fourier expansion of Hammerstein models for nonlinear acoustic system identification," in *Proc. ICASSP*, 2011.

[5] A. Stenger, W. Kellermann, and R. Rabenstein, "Adaptation of acoustic echo cancellers incorporating a memoryless nonlinearity," in *Proc. IWAENC*, 1999.

[6] 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.

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

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

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

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

[11] A. Chhetri, A. Surendran, J. W. Stokes, and J. Platt, "Regression-based residual acoustic echo suppression," in *Proc. IWAENC*, 2005.

[12] 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, Sep. 2006.

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

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