LOW COMPLEXITY NLMS FOR MULTIPLE LOUDSPEAKER ACOUSTIC ECHO CANCELLER USING RELATIVE LOUDSPEAKER TRANSFER FUNCTIONS

Ofer Schwartz [1,2] *Emanuël.A.P Habets* [3] *Sharon Gannot* [2]

[1] CEVA-DSP Sound Department, Herzelia, Israel
[2] Bar-Ilan University Faculty of Engineering, Ramat-Gan, Israel
[3] International Audio Laboratories Erlangen*, Germany

ABSTRACT

Speech signals captured by a microphone mounted to a smart soundbar or speaker are inherently contaminated by echos. Modern smart devices are usually characterized by low computational capabilities and low memory resources; in these cases, a low-complexity acoustic echo canceller (AEC) may be preferred even though a tolerable degradation in the cancellation occurs. In principle, devices with multiple loudspeakers need an individual AEC for each loudspeaker because the transfer function (TF) from each loudspeaker to the microphone must be estimated. In this paper, we present an normalized least mean square (NLMS) algorithm for a multi-loudspeaker case using relative loudspeaker transfer functions (RLTFs). In each iteration, the RLTFs between each loudspeaker and the reference loudspeaker are estimated first, and then the primary TF between the reference loudspeaker and the microphone. Assuming loudspeakers that are close to each other, the RLTFs can be estimated using fewer coefficients w.r.t. the primary TF, yielding a reduction of 3:4 in computational complexity and 1:2 in memory usage. The algorithm is evaluated using both simulated and real room impulse responses (RIRs) of two loudspeakers with a reverberation time set to 0.3 s and several distances between the loudspeakers.

1. INTRODUCTION

Echo cancellation is an important task in home environments where modern smart platforms both generate and listen to audio signals. The general case of multiple microphones and multiple loudspeakers may require high computational capabilities and memory resources, which do not necessarily exist in low-power devices.

The normalized least mean square (NLMS) is the "workhorse" for system identification of echo acoustic paths [1]. Alternative acoustic echo canceller (AEC) algorithms using either recursive least squares (RLS) [2], fast RLS [3], or Kalman filtering [4] require high computational resources. Furthermore, in a multi-microphone and multi-loudspeaker case [5–7], multiple acoustic paths must be identified for each combination of microphone and loudspeaker, which can lead to high computational complexity and memory usage even while using the baseline NLMS.

In [8, 9], the complexity of multi-microphone AECs is reduced using relative echo transfer functions (RETFs). A primary echo signal is estimated using the NLMS, and is then used to compute the secondary acoustic echoes. Using relative transfer functions (RTFs) [10–13], the relation in the frequency domain between the primary and secondary echo signals are estimated and employed. Given closely-spaced microphones, the advantage offered by this approach resides in the fact that RETFs can be modeled using fewer coefficients than the transfer function (TF) between the loudspeaker and the primary microphone. Consequently, the computational complexity can be reduced at the cost of a moderate loss in performance.

In this paper, the concept of using RTF for complexity reduction is adapted to the multi-loudspeaker case (which is an extension of the classic stereophonic case [14–16]) by defining the RTF between two loudspeakers. Namely, an acoustic path between each loudspeaker and microphone are decomposed into two acoustic paths: 1) a virtual path between the loudspeaker and the primary loudspeaker and 2) the primary path between the primary loudspeaker and the microphone. Assuming closely spaced loudspeakers (which is frequently the case in smart soundbars or smart speakers), the first path may be modeled with short time-length w.r.t. the primary path. Thus, addressing the multi-loudspeaker AEC, only a single long TF (the primary TF) and multiple short TFs (the relative loudspeaker transfer functions (RLTFs)) needs to be identified, which reduces the complexity.

2. PROBLEM FORMULATION

Smart speaker devices usually consist of a microphone, multiple loudspeakers, and a centralized processor. The device plays the reference signals using the loudspeakers and simultaneously receives ambient sounds via the microphone. As the device located in enclosure, the ambient sounds consist of reverberated speaker commands, echo signals played by the loudspeakers, and ambient noise. The ultimate goal is to enhance the speaker commands with respect to the echo signals and noise (in this paper the focus is on echo reduction only). The problem formulation is stated in the short-time Fourier transform (STFT) domain, where ℓ and k denote the time-frame index and the frequency bin index, respectively. The signal observed at the microphone is modeled by

$$y(\ell, k) = d(\ell, k) + \sum_{i=1}^{M} z_i(\ell, k) + v(\ell, k), \tag{1}$$

where $d(\ell, k)$ is the desired speech, $z_i(\ell, k)$ is the echo signal from the i-th loudspeaker, and $v(\ell, k)$ is the ambient noise. The echo signals should be entirely cancelled, including their reverberant components. Thus, the echo signals should be modelled using long (w.r.t. the time axis) TFs from the loudspeakers to the microphone. Usually, a short STFT window (e.g. 32 msec) is used for memory and computational reasons. The convolutive transfer function (CTF) model [13, 17] models long impulse responses via a convolution

*The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

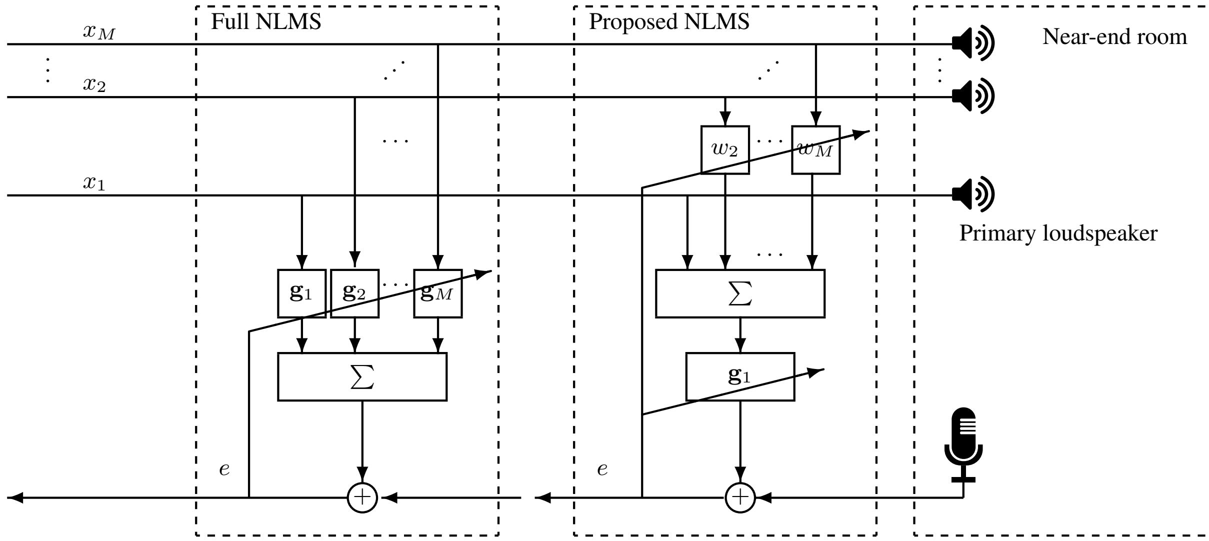


Fig. 1. Diagram of the full NLMS and the proposed RLTF-based NLMS.

operation in the STFT domain, which enables it to model systems that are longer than the STFT window analysis length. Using the CTF model, the echo signals are modelled by the reference signals (played by the loudspeakers) convoluted in the STFT domain with the CTF coefficients that represent the acoustical systems from the loudspeakers to the microphone:

$$z_j(\ell,k) = \sum_{\bar{\ell}=0}^{L-1} g_j(\bar{\ell},k)x_j(\ell-\bar{\ell},k) = \mathbf{g}_j^H(k)\mathbf{x}_j(\ell,k), \quad (2)$$

where $j \in \{1,\ldots,M\}$, the vector $\mathbf{g}_j(k)$ consists of the CTF coefficients, and the vector $\mathbf{x}_j(\ell,k)$ consists of the corresponding time-frames of a known reference signal:

$$\mathbf{g}_j(k) = \left[\begin{array}{c} g_j(0,k), g_j(1,k),\ldots,g_j(L-1,k) \end{array}\right]^{\mathrm{T}}, \quad (3)$$

$$\mathbf{x}_j(\ell,k) = \left[\begin{array}{c} x_j(\ell,k), x_j(\ell-1,k),\ldots,x_j(\ell-L+1,k) \end{array}\right]^{\mathrm{T}} \quad (4)$$

and L is the total number of CTF coefficients.

The goal of this paper is to to cancel the echo signals $z_i(\ell,k)$ and recover the desired speech signal $d(\ell,k)$.

3. NLMS-BASED MULTI-SPEAKER AEC

In the following sections, the frequency index k and frame index ℓ are omitted for brevity whenever no ambiguity arises. First, the full NLMS solution is derived and then the proposed RLTF-based NLMS solution.

3.1. Full NLMS solution

Denote $\mathbf{x} = \left[\mathbf{x}_1^T,\ldots,\mathbf{x}_M^T\right]^T$ and $\mathbf{g} = [\mathbf{g}_1^T,\ldots,\mathbf{g}_M^T]^T$. The general error-signal can be expressed by

$$e[\mathbf{g}] = y - \mathbf{g}^H\mathbf{x}. \quad (5)$$

The NLMS solution for this problem is:

$$\mathbf{g}(\ell) = \mathbf{g}(\ell-1) + \alpha\mathbf{x}, \quad (6)$$

where $\alpha \equiv \frac{\mu_{\mathbf{g}}}{||\mathbf{x}||^2+\epsilon_{\mathbf{g}}} e^*\left[\mathbf{g}(\ell-1)\right]$, $\mu_{\mathbf{g}}$ is the step size and $\epsilon_{\mathbf{g}}$ is regulation factor. While a variable step-size is usually used [18, 19], the derivation for it is beyond the scope of this paper.

3.2. RLTF-based NLMS solution

In low-power platforms (such as the CEVA platform used in Sec. 5.3), the computational complexity and memory usage is critical when implementing any algorithm. The majority of the computational complexity is related to the length of the TFs L and the number of loudspeakers M. Adopting the RETF principle from [8], we propose using RLTF to reduce the computations even in the multi-loudspeaker case. In cases where the loudspeakers are sufficiently closely-spaced, modelling M long TFs for the loudspeakers may be avoided by using a single long TF for one of the loudspeakers and $M-1$ short RLTFs. In practice, without loss of generality, system $\mathbf{g}_i \quad \forall i = 1,\ldots,M$ can be modelled by two acoustic paths, the first from the i-th loudspeaker to the primary loudspeaker and the second from the primary loudspeaker to the microphone. In this case, the first path can be modelled by a shorter TF than the second path. In this paper, the first system is modeled using a single coefficient; however, a general model can consist of Q coefficients with $Q < L$.

Denote w_i as the RLTF. Accordingly, the error-signal can be redefined using only the system $\mathbf{g}_1$ and the corresponding RLTF:

$$e[\mathbf{g}_1,\mathbf{w}] = y - \mathbf{g}_1^H\left(\mathbf{x}_1 + \sum_{i=2}^{M} w_i^*\mathbf{x}_i\right) \quad (7)$$

The systems $\mathbf{g}_1$ and $\mathbf{w} = [w_2,\ldots,w_M]^T$ can be estimated separately by

$$\mathbf{w}(\ell) = \mathbf{w}(\ell-1) + \beta\mathbf{h}, \quad (8)$$

where

$$\beta \equiv \frac{\mu_w}{||\mathbf{h}||^2+\epsilon_w} e^*\left[\mathbf{g}_1(\ell-1),\mathbf{w}(\ell-1)\right] \quad (9a)$$

$$\mathbf{h} \equiv \left[\mathbf{g}_1^H(\ell-1)\mathbf{x}_2,\ldots,\mathbf{g}_1^H(\ell-1)\mathbf{x}_M\right] \quad (9b)$$

and

$$\mathbf{g}_1(\ell) = \mathbf{g}_1(\ell-1) + \gamma\mathbf{f}, \quad (10)$$

with

$$\gamma \equiv \frac{\mu_{\mathbf{g}}}{||\mathbf{f}||^2+\epsilon_{\mathbf{g}}} e^*\left[\mathbf{g}_1(\ell-1),\mathbf{w}(\ell)\right] \quad (11a)$$

$$\mathbf{f} \equiv \mathbf{x}_1 + \sum_{i=2}^{M} w_i^*(\ell)\mathbf{x}_i. \quad (11b)$$

The proposed NLMS and the full NLMS are described in Fig. 1.

Eq.	Operation	Multiplications
(5)	$\mathbf{g}^H\mathbf{x}$	LM
(6)	$\|\|\mathbf{x}\|\|^2$	LM
(6)	$\alpha\mathbf{x}$	LM
Total		3LM

Table 1. Multiplications for the full NLMS.

Eq.	Operation	Multiplications
(7)	$\mathbf{g}_1^H\mathbf{x}_i$	LM
(9a)	$\|\|\mathbf{h}\|\|^2$	M-1
(8)	$\beta\mathbf{h}$	M-1
(11b)	$\sum_{i=2}^{M} w_i^*(\ell)\mathbf{x}_i$	L(M-1)
(11a)	$\|\|\mathbf{f}\|\|^2$	L
(10)	$\gamma\mathbf{f}$	L
Total		2LM + L + 2M-2

Table 2. Multiplications for the proposed NLMS.

4. COMPUTATIONAL COMPLEXITY AND MEMORY USAGE

In this section, the two NLMS algorithms are compared in terms of computational complexity and memory usage. Only the multiplications and memory usage that depend on L or M are summed. The multiplications for the full NLMS are summarized in Table 1, and the multiplications for the proposed NLMS are summarized in Table 2. The multiplications for both of the NLMS options are $\mathcal{O}(LM)$. When there is a high number of coefficients L and loudspeakers M, the multiplication ratio between the approaches equals 2 : 3.

As for memory consumption, both algorithms require ML memory cells for $\mathbf{x}$. However, the full NLMS requires excess ML memory cells for $\mathbf{g}$, and the proposed NLMS requires $L + M - 1$ memory cells for $\mathbf{g_1}$ and $\mathbf{w}$. Thus, for a higher number of coefficients and loudspeakers, the memory usage ratio approaches 1 : 2.

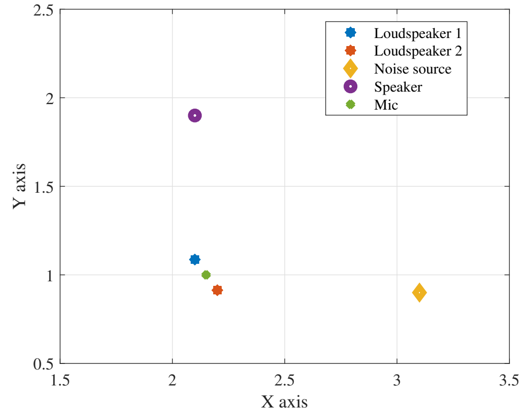
5. PERFORMANCE EVALUATION

The performance of the proposed algorithms is evaluated in terms of output echo return loss enhancement (ERLE) and perceptual evaluation of speech quality (PESQ) [20] scores. The experiments consist of speech signals and reference music signals convoluted with corresponding 1) simulated room impulse responses (RIRs) and 2) real RIRs extracted by the CEVA platform. Using the simulated scenario, the influence of the distance between the loudspeakers on the performance of the proposed algorithm is tested. The proposed algorithm is compared to the full NLMS elaborated in Sec. 3.

5.1. Setup for the experiments

Four loudspeakers (which model the speaker, two echo sources and noise source) were received by a single microphone. Anechoic speech, music signals (left and right references) and pink noise were convoluted with RIRs generated by an open-source RIRs simulator[1] or, alternatively with real-life RIRs extracted using CEVA "smart and connected" development platform.[2] The RIRs were extracted by playing chirp signals from the various positions reported in Fig. 2 (a), and then extracting the corresponding RIRs using frequency division. The distance between the loudspeakers was 0.25 m.

The sampling frequency was 16 kHz and the frame length of the STFT was 64 ms (1024 sample lengths of the analysis window) with 16 ms between successive time frames (i.e., 75% overlap). The algorithms were applied in speech absence segments using a perfect voice activity detector (VAD). In practice, segments with double-talk (namely, when the echo and the near-end speaker are simultaneously active) should be detected and the echo-path identification should be freezed. This can be implemented by the step-size adaptation [18, 19, 21, 22]. The length of the primary TF was set to $L = 8$ for $T_{60} = 0.3$ and to $L = 16$ for $T_{60} = 0.6$. Note that the proposed AEC is implemented in the time-frequency domain, hence N taps cover 64 ms for the first tap and additional 16 ms for all subsequent $N-1$ taps. The two step-sizes $\mu_{\mathbf{g}}$ and μ_w were set to 0.5 and 0.005, respectively, and the regulation factors $\epsilon_{\mathbf{g}}, \epsilon_w$ were set to 1 and 10^{-4}. The values were empirically adjusted to achieve the best



(a)

(b)

Fig. 2. (a) Geometric setup for simulated experiment (b) CEVA "smart and connected" development platform with two connected loudspeakers and single loudspeaker as the desired speaker.

[1]The RIRs simulator can be freely downloaded from https://github.com/ehabets/RIR-Generator

[2]For more details see https://www.ceva-dsp.com/product/smart-connected-development-platform/

Alg. \ Distance (m)	0.05	0.1	0.2	0.3	0.5	1
Full NLMS	42.98	42.92	42.91	42.85	42.50	42.03
Proposed NLMS	42.03	41.54	40.62	39.72	39.15	37.52
Input	1.08	1.09	1.08	1.08	1.08	1.09
Full NLMS	3.00	3.02	3.05	3.08	3.09	3.06
Proposed NLMS	2.82	2.72	2.64	2.55	2.51	2.42

Table 3. ERLE results (top) and PESQ results (bottom) with various distances between the loudspeakers for $T_{60} = 0.3$ and 8 taps

Alg. \ Distance (m)	0.05	0.1	0.2	0.3	0.5	1
Full NLMS	36.56	36.45	36.24	36.03	35.64	34.68
Proposed NLMS	38.74	38.09	36.56	35.44	34.83	32.45
Input	1.11	1.11	1.10	1.10	1.09	1.10
Full NLMS	3.15	3.18	3.22	3.22	3.19	3.17
Proposed NLMS	3.07	2.93	2.79	2.69	2.72	2.55

Table 4. ERLE results (top) and PESQ results (bottom) with various distances between the loudspeakers for $T_{60} = 0.6$ and 16 taps.

results.

The scores were computed by averaging the results obtained using 100 signals, each of which 28 sec long, evenly distributed between male and female speakers. The input signal to echo ratio (SER) was -5 dB and is defined by

$$\mathrm{SER}_{\mathrm{in}} = 10 \log \left(\frac{\sum_{k,\ell} |d|^2}{\sum_{k,\ell} |z_1 + z_2|^2} \right). \quad (12)$$

The noise source was added with a 40 dB signal-to-noise ratio (SNR) (w.r.t. the speech signal). The ERLE results were calculated using the ratio between the overall echo and the estimated error, i.e.,

$$\mathrm{ERLE} = 10 \log \left(\frac{\sum_{k,\ell} |z_1 + z_2|^2}{\sum_{k,\ell} |z_1 + z_2 - \hat{z}_1 - \hat{z}_2|^2} \right). \quad (13)$$

PESQ results were obtained using the desired speech d and its estimate e.

5.2. Experiment using simulated RIRs

The reverberation time T_{60} was set to 0.3 and 0.6 sec. An illustration of the geometric setup is given in Fig. 2 (a). In Table 4, ERLE and PESQ results are presented for several distances between the loudspeakers. It can be observed that the full NLMS algorithm achieves the best results. However, the proposed RLTF-based NLMS achieves high scores especially for small distances (up to 10 cm). The performance is indeed degraded for large distances. It is therefore recommended to use the proposed algorithm only for small distances between the loudspeakers (< 0.1 meter).

5.3. Experiment using RIRs extracted by real recordings

Two DELL loudspeakers were used (with 20cm inter distance) while the desired speech was played using another isolated loudspeaker 1m from the DSP platform. The geometric setup can be seen in Fig. 2 (b). Sonogram examples of the input signal, the sum of the reference signals, the full NLMS output, and the proposed NLMS output are depicted in Fig. 3. The near-end speech was a set of 157 trigger-words. The various signals were tested by WhisPro, a recursive neural network based speech recognition software package provided by CEVA.[3] The detection results were 38.8% for the observed signal, 91% for the full NLMS algorithm and 77% for the RLTF-based algorithm. It can be seen by the sonograms and verified by the detection results that both of the algorithms cancel the echo signal while maintaining the speech signal. However, the full NLMS achieves better cancellation (albeit with more computational complexity).

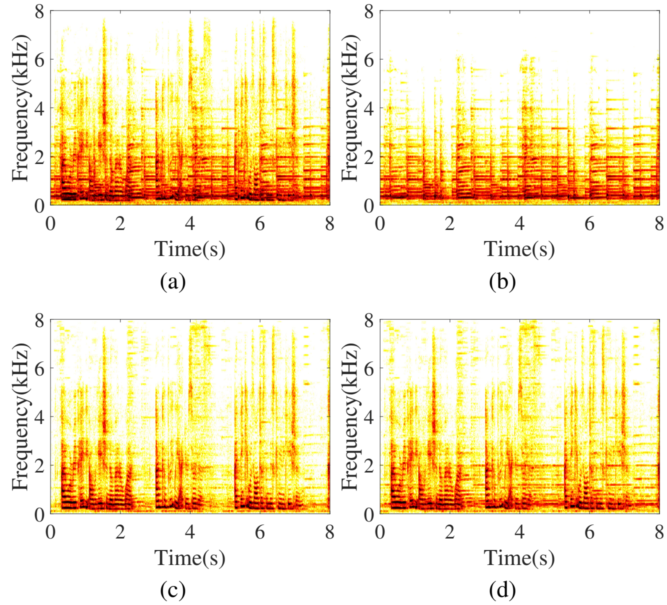


Fig. 3. Example sonograms. (a) Observed signal. (b) Left reference signal. (c) Full NLMS output. (d) Proposed NLMS output.

6. CONCLUSIONS

In this paper, an NLMS algorithm for a multi-loudspeaker case using RLTFs is presented. The RLTFs between each loudspeaker and the reference loudspeaker is estimated first, and then the primary TF between the reference loudspeaker and the microphone. Assuming closely-spaced loudspeakers, the RLTFs can be estimated using fewer coefficients w.r.t. the primary TF, yielding a reduction of 2:3 in computational complexity and 1:2 in memory usage. The algorithm is evaluated using simulated and real RIRs of two loudspeakers with a reverberation time set to 0.3 s and several distances between the loudspeakers. Compared to the full stereo-NLMS AEC and using same tap length for the primary-path, the proposed RLTF based AEC achieves less cancellation while using less computations. For small inter-loudspeakers distances the results are adjacent.

7. REFERENCES

[1] Eberhard Hänsler and Gerhard Schmidt, *Acoustic echo and noise control: A practical approach*, vol. 40, John Wiley & Sons, 2005.

[2] Camelia Elisei-Iliescu, Constantin Paleologu, Cristian Stanciu, Cristian Anghel, Silviu Ciochina, and Jacob Benesty, “A practical overview of recursive least-squares algorithms for echo

[3] Further details can be found in `https://www.ceva-dsp.com/product/ceva-whispro/`

cancellation," in *The Sixteenth International Conference on Networks (ICN)*, 2017, pp. 87–91.

[3] Jacob Benesty, Fabrice Amand, André Gilloire, and Yves Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *IEEE Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995, vol. 5, pp. 3099–3102.

[4] Constantin Paleologu, Jacob Benesty, and Silviu Ciochină, "Study of the general kalman filter for echo cancellation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 8, pp. 1539–1549, 2013.

[5] Kentaro Koga, Tetsuya Takiguchi, and Yasuo Ariki, "Echo canceller for multi-loudspeakers based on maximum likelihood using an acoustic model," in *Asia-Pacific Signal and Information Processing Association*, 2009, pp. 246–249.

[6] Herbert Buchner, "Acoustic echo cancellation for multiple reproduction channels: From first principles to real-time solutions," in *ITG Conference on Voice Communication*, 2008.

[7] Walter Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, vol. 1, pp. 219–222.

[8] María Luis Valero and Emanuël A.P. Habets, "Multi-microphone acoustic echo cancellation using relative echo transfer functions," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 229–233.

[9] María Luis Valero and Emanuël A.P. Habets, "Low-complexity multi-microphone acoustic echo control in the short-time fourier transform domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 595–609, 2018.

[10] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[11] Israel Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[12] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[13] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 546–555, 2009.

[14] Tomas Gänsler and Jacob Benesty, "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," *International Journal of adaptive control and signal processing*, vol. 14, no. 6, pp. 565–586, 2000.

[15] Shoji Makino, "Stereophonic acoustic echo cancellation: An overview and recent solutions," *Acoustical Science and Technology*, vol. 22, no. 5, pp. 325–333, 2001.

[16] Jacob Benesty, Dennis R Morgan, and Man Mohan Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE transactions on speech and audio processing*, vol. 6, no. 2, pp. 156–165, 1998.

[17] Yekutiel Avargel and Israel Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[18] Constantin Paleologu, Jacob Benesty, Steven L Grant, and Christopher Osterwise, "Variable step-size NLMS algorithms designed for echo cancellation," in *Proc. of the 43rd IEEE Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 633–637.

[19] Jean-Marc Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1030–1034, 2007.

[20] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

[21] Jacob Benesty, Hernan Rey, Leonardo Rey Vega, and Sara Tressens, "A nonparametric vss nlms algorithm," *IEEE Signal Processing Letters*, vol. 13, no. 10, pp. 581–584, 2006.

[22] Constantin Paleologu, Silviu Ciochină, Jacob Benesty, and Steven L Grant, "An overview on optimized nlms algorithms for acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 97, 2015.