Double-Talk Robust Acoustic Echo Cancellation Using Partition Block Frequency-Domain Adaptive Filtering

Clément Gaultier
Orange Labs
Cesson-Sévigné, France

Alexandre Guérin
Orange Labs
Cesson-Sévigné, France

Grégory Pallone
Orange Labs
Cesson-Sévigné, France

Marc Emerit
Orange Labs
Cesson-Sévigné, France

***Abstract*—This work aims at introducing a new acoustic echo cancellation algorithm robust to double-talk situation. We propose a method which combines partition-block frequency domain adaptive filtering (PBFDAF) and best linear unbiased estimation (BLUE) through local signal characteristics recursive estimation. We report that our method outperforms an Acoustic Echo Cancellation (AEC) baseline using BLUE for Echo Return Loss Enhancement (ERLE) and show its usefulness when used in an automated speech recognition pipeline. Improvements are obtained for situations featuring either continuous or discontinuous echo signals.**

***Index Terms*—Acoustic Echo Cancellation, Double-Talk, BLUE**

I. Introduction

The increasing demand for hands-free telephony and more recently the wide use of teleconferencing systems motivated the long standing interest for the acoustic echo cancellation (AEC) problem. In a classical AEC scenario (fig. 1), a reference signal $x(t)$ is reproduced in a room. In this room a microphone aims at recording a local signal $s(t)$. An acoustic path $w(t)$ exists between the microphone and the sound reproduction system such that the microphone not only records the local signal but also an undesired echo $d(t)$ originating from the reference signal. The main goal of AEC is to retrieve the local signal by subtracting the echo signal estimate from the microphone. In order to form the acoustic echo estimate, the procedure needs an estimate of the acoustic path. While it is often time dependent, the AEC usually relies on adaptive filtering to recursively estimate $w(t)$.

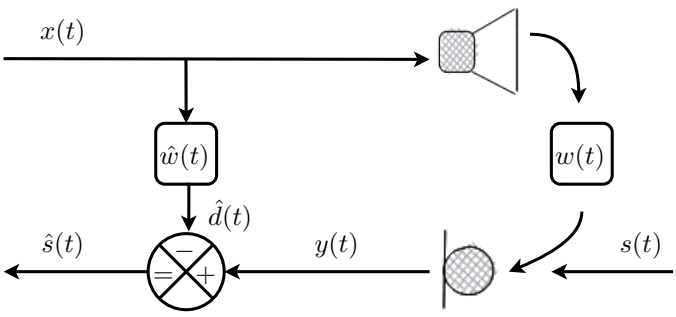


Fig. 1: AEC pipeline

Traditional time-domain adaptive filtering methods (Normalized Least Mean Squares - NLMS [1] or Recursive Least Squares - RLS) as well as frequency-domain NLMS [2] approaches identification performance drops in the presence of both local signal $s(t)$ and echo also called "double-talk" (DT) periods. To tackle this problem, a full body of work equipped adaptive filters with double-talk detectors (DTDs) [3], [4]. The purpose of such DTD, specifically designed for AEC applications, is to slow down or even freeze the adaptive filtering update during DT periods. However, some situations (continuous or fast bursting double-talk) make the DTDs unusable, calling the need for DT robust approaches without any DTD.

Several other approaches were developed to ensure robustness to DT situations. Some of them rely on variable step size (VSS) to control adaption during DT [5]. Other use a minimum-variance linear estimation solution rather than a more common least squares approach to estimate the acoustic path [6], [7]. The latter, also known as Best Linear Unbiased Estimation (BLUE) [8], depends on the local signal properties to produce an optimal acoustic path estimate during double-talk.

A. Contributions and outline

In this paper we aim at developing a recursive adaptive filtering algorithm allowing acoustic path estimation $\hat{\boldsymbol{w}}$ robust to either continuous or discontinuous double-talk situations without the need of any double-talk detector. We propose a double-talk robust AEC method based on a twofold strategy: a partition-block frequency domain adaptive filtering together with a minimum-variance linear estimation (BLUE). Section II describes our notations and model. The proposed method is presented in section III. More precisely, we describe in section III-B local signal frequency characteristics recursive short-time estimation. The approach is validated on real audio data in section IV. Conclusions and future directions are listed in section V.

II. Notations and Model

A. Notations

In the following, lower-case sans serif font (i) denotes an integer. Lower-case bold font ($\mathbf{v}, \boldsymbol{v}$) expresses vectors and

upper-case ($\mathbf{V}, \boldsymbol{V}$) matrices. * superscript stands for complex conjugate and $^{\mathsf{H}}$ for hermitian transpose. $\hat{\mathbf{v}}$ (resp. $\hat{\mathbf{V}}$) denotes an estimate of $\mathbf{v}$ (resp. $\mathbf{V}$). $\mathbf{v}[\mathsf{i}]$ is the i^{th} component of a vector $\mathbf{v}$. Other notations will be disambiguated in the text.

B. Signal Model

We assume in the following that acoustic indoor propagation ($w(t)$) can be modeled with finite impulse response $w(t) = \boldsymbol{w} \in \mathbb{R}^{\mathsf{N}}$. Hence, the echo ($d(t)$) signal originating from a discrete reference signal $x(t)$ reproduced by a loudspeaker inside a room (and encompassing all the reflections) can be expressed by:

$$d(t) = x(t) \star w(t), \tag{1}$$

with $\star$ denoting the convolution. Therefore, we consider in such a setting a microphone signal modeled as:

$$y(t) = x(t) \star w(t) + s(t), \tag{2}$$

with $s(t)$ representing some signal of interest sometimes called "local signal" in a communication setup, as opposed to the reference signal $x(t)$ also called "distant signal". Right part of fig. 1 illustrates microphone signal and acoustic path models.

III. Double-Talk robust PBFDLMS algorithm

The goal is to estimate the signal $s(t)$ by removing the echo from the microphone signal $y(t)$ (*i.e.* performing AEC). For that, we propose an algorithm working on a frame-base manner based on the Partition Block Frequency-Domain Adaptive Filtering (PBFDAF) method [9] specifically designed to account for double-talk situations. We build our approach on PBFDAF allowing for long acoustic path identification and reduced latency.

A. Least Mean Squares PBFDAF framework

We observe a single channel time-domain audio signal $y(t)$. A M-samples long frame of such a signal is denoted $\boldsymbol{y} \in \mathbb{R}^{\mathsf{M}}$. Similarly, we denote $\boldsymbol{x} \in \mathbb{R}^{\mathsf{L}}$ a frame with L consecutive samples from the reference signal $x(t)$. $\boldsymbol{w} \in \mathbb{R}^{\mathsf{N}}$ is the time-domain finite impulse response modeling the underlying acoustic path $w(t)$. $\mathbf{F} \in \mathbb{C}^{\mathsf{M}\times\mathsf{L}}$ is a matrix encompassing a possibly redundant discrete frequency-domain transform (*i.e.* Discrete Fourier Transform (DFT)). We form $\mathbf{W} \in \mathbb{C}^{\mathsf{M}\times\mathsf{P}}$ a time-frequency representation of $\boldsymbol{w}$ such that:

$$\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_{\mathsf{P}}], \mathbf{w}_{\mathsf{p}} \in \mathbb{C}^{\mathsf{M}}, \mathbf{w}_{\mathsf{p}} \overset{\forall \mathsf{p}}{=} \mathbf{F}\boldsymbol{w}_{\mathsf{p}}, \tag{3}$$

with $\boldsymbol{w}_{\mathsf{p}} \in \mathbb{R}^{\mathsf{L}}$ a partition of $\boldsymbol{w}$ and $\mathsf{P} \times \mathsf{L} \leq \mathsf{N}$. In a similar way, we consider $\mathbf{X} \in \mathbb{C}^{\mathsf{M}\times\mathsf{P}}$ a circulant matrix gathering the frequency transforms of the last P overlapping frames from the reference signal $x(t)$ such that:

$$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_{\mathsf{P}}], \mathbf{x}_{\mathsf{p}} \in \mathbb{C}^{\mathsf{M}}, \mathbf{x}_{\mathsf{p}} \overset{\forall \mathsf{p}}{=} \mathbf{F}\boldsymbol{x}_{\mathsf{p}}, \tag{4}$$

with $\boldsymbol{x}_{\mathsf{p}} \in \mathbb{R}^{\mathsf{L}}$ a time frame from $x(t)$. $\boldsymbol{y} \in \mathbb{R}^{\mathsf{L}}$ contains the L most recent samples extracted from the microphone signal $y(t)$. We denote $\mathbf{y} \in \mathbb{C}^{\mathsf{M}}$ such that:

$$\mathbf{y} = \mathbf{F}\begin{bmatrix}\mathbf{0}_{\mathsf{M}-\mathsf{L}} \\ \boldsymbol{y}\end{bmatrix}. \tag{5}$$

k is the time-frame index. Hence, the algorithm takes L new samples from the k-th overlapping frame of $x(t)$ and produces L filtered samples $\hat{\boldsymbol{s}}$ along with an estimate of the acoustic path $\hat{\mathbf{W}}$. Equations eqs. (6) to (8) describe the PBFDAF steps using the overlap-save technique to prevent approximations introduced by circular operations performed in the frequency domain.

$$\hat{\boldsymbol{s}}^{(\mathsf{k})} = \begin{bmatrix}\mathbf{0}_{\mathsf{M}-\mathsf{L}} \\ \boldsymbol{y}^{(\mathsf{k})}\end{bmatrix} - \begin{bmatrix}\mathbf{0}_{\mathsf{M}-\mathsf{L}} \\ \mathbf{1}_{\mathsf{L}}\end{bmatrix} \mathbf{F}^{\mathsf{H}} \sum_{\mathsf{p}=1}^{\mathsf{p}=\mathsf{P}} \mathbf{w}_{\mathsf{p}}^{(\mathsf{k})} \circ \mathbf{x}_{\mathsf{p}}^{*(\mathsf{k})} \tag{6}$$

$$\Delta\mathbf{w}_{\mathsf{p}}^{(\mathsf{k})} \overset{\forall \mathsf{p}}{=} \mathbf{G}\boldsymbol{\Lambda}_{\mathsf{p}}^{(\mathsf{k})} \circ \mathbf{x}_{\mathsf{p}}^{*(\mathsf{k})} \circ \mathbf{F}\hat{\boldsymbol{s}}^{(\mathsf{k})} \tag{7}$$

$$\hat{\mathbf{w}}_{\mathsf{p}}^{(\mathsf{k}+1)} \overset{\forall \mathsf{p}}{=} \hat{\mathbf{w}}_{\mathsf{p}}^{(\mathsf{k})} + \Delta\mathbf{w}_{\mathsf{p}}^{(\mathsf{k})} \tag{8}$$

where $\circ$ denote the Hadamard product, $\boldsymbol{\Lambda}_{\mathsf{p}}^{(\mathsf{k})} \in \mathbb{C}^{\mathsf{M}}$ is the frequency dependent normalization term applied to the filter update. Computation details are given through section III-B. $\mathbf{G} \in \mathbb{C}^{\mathsf{M}\times\mathsf{M}}$ defined below allows for either a "constrained" or "unconstrained" filter update [9], [10].

Unconstrained update	Constrained update
$\mathbf{G} = \mathbf{I}_{\mathsf{M}}$	$\mathbf{G} = \mathbf{F}\mathbf{F}^{\mathsf{H}}$

B. Double-talk robust filter update

We know from [11] that contrarily to the classical least squares approach, the Best Linear Unbiased Estimation (BLUE) gives optimal acoustic path estimate during double-talk periods. As used in [12], BLUE of acoustic path (eq. (9)) depends on the characteristics of the local signal $s(t)$ (more precisely $\boldsymbol{\Gamma}_{\mathbf{s}}$ its power spectral density). However these are unknown and often highly time-dependent [6]. $\hat{\mathbf{W}}$ estimation of the underlying acoustic path is given by minimizing the following:

$$\left(\mathbf{y} - \textstyle\sum_{\mathsf{p}=1}^{\mathsf{p}=\mathsf{P}} \mathbf{w}_{\mathsf{p}}^{(\mathsf{k})} \circ \mathbf{x}_{\mathsf{p}}^{*(\mathsf{k})}\right)^{\mathsf{T}} \left(\gamma\boldsymbol{\Gamma}_{\mathbf{s}} + \mathbf{x}_{\mathsf{p}}^{\mathsf{T}}\mathbf{x}_{\mathsf{p}}\right)^{-1} \left(\mathbf{y} - \textstyle\sum_{\mathsf{p}=1}^{\mathsf{p}=\mathsf{P}} \mathbf{w}_{\mathsf{p}}^{(\mathsf{k})} \circ \mathbf{x}_{\mathsf{p}}^{*(\mathsf{k})}\right) \tag{9}$$

In the light of "regularized" Best Linear Unbiased Estimation (BLUE) [12], we seek here a solution which produces minimum residual echo and account for double-talk situations without requiring any double-talk detector (DTD) or voice activity detection (VAD) step. We expect that estimating the short-time local signal ($s(t)$) properties simultaneously with the acoustic path can improve AEC in double-talk situations. We base ou approach on a specific filter update (eq. (7)) normalization $\boldsymbol{\Lambda}_{\mathsf{p}}^{(\mathsf{k})}$. In order to derive $\boldsymbol{\Lambda}_{\mathsf{p}}^{(\mathsf{k})}$, we first perform the following steps:

- Microphone and reference signals Power Spectral Densities and Cross Power Spectral Densities estimation,
- Echo to Signal Ratio estimation,
- Local signal $s(t)$ Power Spectral Density estimation,

detailed hereafter.

a) Power Spectral Densities (PSD) estimations:
We consider $\mathbf{\Gamma_{x_p}} \in \mathbb{R}^{\mathsf{M}}$ (resp. $\mathbf{\Gamma_y} \in \mathbb{R}^{\mathsf{M}}$) the PSD of the p^{th} frame of the signal $x(t)$ (resp. of $y(t)$). Similarly, $\mathbf{\Gamma_{s_p}} \in \mathbb{R}^{\mathsf{M}}$ is the PSD of the p^{th} frame of the signal $s(t)$. We also denote $\mathbf{\Gamma_{yx_p}}$ the cross PSD for that corresponding frame. PSD and cross PSD estimates are then expressed as follow:

$$\hat{\mathbf{\Gamma}}_{\mathbf{x_p}}^{(\mathsf{k})}[\mathsf{m}] \overset{\forall \mathsf{p},\mathsf{m}}{=} \alpha \hat{\mathbf{\Gamma}}_{\mathbf{x_p}}^{(\mathsf{k}-1)}[\mathsf{m}] + (1-\alpha)|\mathbf{x}_{\mathsf{p}}^{(\mathsf{k})}[\mathsf{m}]|^2, \tag{10}$$

$$\hat{\mathbf{\Gamma}}_{\mathbf{y}}^{(\mathsf{k})}[\mathsf{m}] \overset{\forall \mathsf{m}}{=} \eta \hat{\mathbf{\Gamma}}_{\mathbf{y}}^{(\mathsf{k}-1)}[\mathsf{m}] + (1-\eta)|\mathbf{y}^{(\mathsf{k})}[\mathsf{m}]|^2. \tag{11}$$

For better readability we drop the frequency index m in the remaining equations, however these hold for each partition p and each frequency component m:

$$\hat{\mathbf{\Gamma}}_{\mathbf{yx_p}}^{(\mathsf{k})} \overset{\forall \mathsf{p}}{=} \begin{cases} \xi \hat{\mathbf{\Gamma}}_{\mathbf{yx_p}}^{(\mathsf{k}-1)} + (1-\xi)|\mathbf{yx_p}|^2 & \text{if} \quad \hat{\mathbf{\Gamma}}_{\mathbf{yx_p}}^{(\mathsf{k}-1)} \leq |\mathbf{yx_p}|^2, \\ \left(\delta \sqrt{\hat{\mathbf{\Gamma}}_{\mathbf{yx_p}}^{(\mathsf{k}-1)}} + (1-\delta)|\mathbf{yx_p}|\right)^2 & \text{otherwise,} \end{cases} \tag{12}$$

with α, η, ξ and $\delta \in (0,1)$.

b) Instantaneous Echo to Signal Ratio (ESR) estimation:

We consider $\mathbf{\Pi_p} \in \mathbb{R}^{\mathsf{M}}$ the instantaneous ESR for partition p. We derive its estimate as:

$$\hat{\mathbf{\Pi}}_{\mathsf{p}}^{(\mathsf{k})} \overset{\forall \mathsf{p}}{=} \beta \frac{\hat{\mathbf{\Gamma}}_{\mathbf{y}}}{\hat{\mathbf{\Gamma}}_{\mathbf{s_p}}^{(\mathsf{k}-1)}} \circ \frac{\hat{\mathbf{\Pi}}_{\mathsf{p}}^{(\mathsf{k}-1)}}{1+\hat{\mathbf{\Pi}}_{\mathsf{p}}^{(\mathsf{k}-1)}} + (1-\beta)\left(\frac{\hat{\mathbf{\Gamma}}_{\mathbf{yx_p}}^{(\mathsf{k})}}{\hat{\mathbf{\Gamma}}_{\mathbf{x_p}}^{(\mathsf{k})}} \circ \frac{1}{\hat{\mathbf{\Gamma}}_{\mathbf{s_p}}^{(\mathsf{k}-1)}}\right), \tag{13}$$

and $\beta \in (0,1)$. This step, relying on short-time power spectral densities estimations links to decision-directed Signal-to-Noise Ratio estimation approach [13].

c) Local signal PSD Estimation:
From the estimates of the microphone PSD and the instantaneous ESR, we write the estimate PSD of the local signal as:

$$\hat{\mathbf{\Gamma}}_{\mathbf{s_p}}^{(\mathsf{k})} \overset{\forall \mathsf{p}}{=} \begin{cases} \frac{\hat{\mathbf{\Gamma}}_{\mathbf{y}}^{(\mathsf{k})}}{1+\hat{\mathbf{\Pi}}_{\mathsf{p}}^{(\mathsf{k})}} & \text{if} \quad \hat{\mathbf{\Pi}}_{\mathsf{p}}^{(\mathsf{k})} \leq \zeta, \\ \hat{\mathbf{\Gamma}}_{\mathbf{s_p}}^{(\mathsf{k}-1)} & \text{otherwise,} \end{cases} \tag{14}$$

with $\zeta \in \mathbb{R}_{>0}$.

Once all the previous estimates are available, we derive a normalization term depending on the partition and the frequency $\mathbf{\Lambda}^{(\mathsf{k})} = [\mathbf{\Lambda}_1^{(\mathsf{k})}, ..., \mathbf{\Lambda}_{\mathsf{p}}^{(\mathsf{k})}, ..., \mathbf{\Lambda}_{\mathsf{P}}^{(\mathsf{k})}] \in \mathbb{R}^{\mathsf{M}\times\mathsf{P}}$ to apply to the filter update step such that:

$$\mathbf{\Lambda}_{\mathsf{p}}^{(\mathsf{k})} = \frac{\mu}{\hat{\mathbf{\Gamma}}_{\mathbf{x_p}}^{(\mathsf{k})} + \gamma \hat{\mathbf{\Gamma}}_{\mathbf{s_p}}^{(\mathsf{k})}}, \tag{15}$$

with $\mu \in (0,1]$ and $\gamma \in \mathbb{R}_{>0}$. When γ is close to 0, the algorithm then recasts as the simple Partition Block Frequency-Domain "Normalized" Least Mean Squares [9].

IV. Experiments

A. Experimental setup

In order to validate the algorithm, we perform experiments on single channel audio data targeting 2 different scenarii. One is dedicated to vocal assistant interactions (someone requesting something while music is playing) referred to as "Vocal Assistant" and another dedicated to interpersonal communication between two persons referred to as "Communication". We choose as the local signal $s(t)$ to recover, speech excerpts from the gender-balanced 120 speakers French corpus BREF [14]. For the "Vocal Assistant" use case, as reference signals $x(t)$ we choose monophonic versions of the RWC Pop dataset [15]. For the "Communication" use case, as reference signals, we also choose speech from the French corpus BREF. Finally, for acoustic paths $w(t)$ we use the MARDY dataset [16]. We artificially generated echo signals randomly picking a filter from the MARDY database and convolving it to a reference signal. Input signals $y(t)$ are obtained by mixing those echo signals with $s(t)$ at 4 different Signal-to-Echo Ratio (SER) {-20dB, -10dB, 0dB, 10dB}. We compare the proposed method with our implementation of a frequency domain AEC baseline [12] able to cope with double-talk situations thanks to BLUE and information on the local signal acquired during low-energy echo period. For the "Communication" use case, vocal activity detection on the reference signal (provided by [17]) allows [12] to estimate local signal PSD. For the "Vocal Assistant" use case, a 3 seconds time period with only local signal at the begining of the sound excerpt allows such an estimation. We also compare with the same baseline method benefiting from ground-truth local signal PSD values (denoted as " [12] w/ Oracle PSD" in the following). Experimental parameters are summarized in table I. Practically, redundant frequency transform ($\mathbf{F} \in \mathbb{C}^{\mathsf{M}\times\mathsf{L}}, \mathsf{M} > \mathsf{L}$) is achieved through zero-padding on time-domain signals.

Signal based measures We compare the performance of the adaptive filtering methods with the Echo Return Loss Enhancement (ERLE) index expressed in dB:

$$ERLE = -10\log\left(\frac{\|\hat{\boldsymbol{s}} - \boldsymbol{s}\|_2^2}{\|\boldsymbol{y} - \boldsymbol{s}\|_2^2}\right). \tag{16}$$

Note that this definition of ERLE uses direct access to the residual echo and undistorted echo.

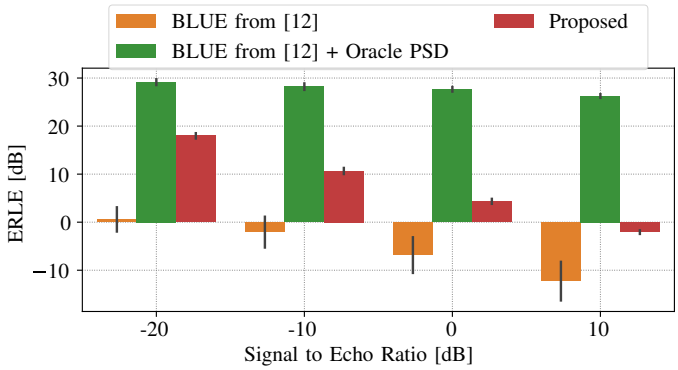
Automated Speech Recognition evaluation Finally we run a comparison on automated speech recognition (ASR) task and report the Word Error Rate (WER) obtained by Cobalt Speech Recognition, developed by Orange Labs for French ASR. It is a Kaldi-based speech-to-text de-coder [18] using a time-delay neural network based acoustic model [19] trained on more than 2000 h of clean and noisy speech, a 1.7-million-word lexicon, and a 5-gram language model trained on 3 billion words.

B. Results

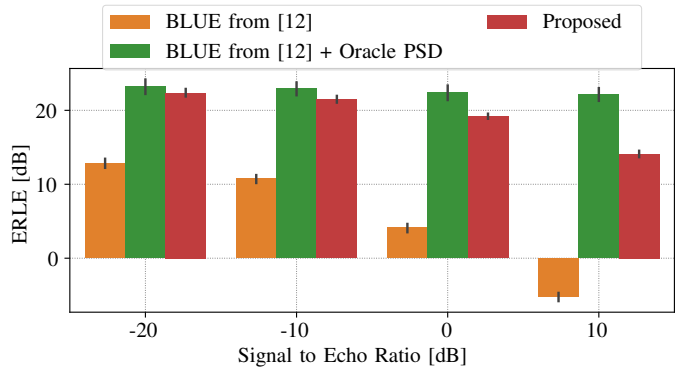
Bar plots presented in this section are average performance results across the 120 tested speakers. Thin black vertical bars denote 95% confidence intervals. Figure 2 shows that for either

TABLE I: Experimental parameters

Parameter	Sampling Frequency	Overlap	Estimated filter size	Number of partitions	μ	Frequency transform	Constrained update	$\alpha, \eta, \xi, \delta$	β	ζ
Value	16 kHz	50 %	192 ms	P = 12	0.5	$\mathbf{F}$ = DFT	$\mathbf{G} = \mathbf{F}^{\mathrm{H}}\mathbf{F}$	0.8	0.98	$1 \cdot 10^{10}$



(a) Communication



(b) Vocal assistant

Fig. 2: Signal to Echo Ratio *vs* Echo Return Loss Enhancement

tested use-cases (Vocal assistant: *i.e.* continuous echo and discontinuous local signal, Communication: *i.e.* discontinuous echo and local signal) the proposed approach outperforms the baseline and is even on par with the method benefiting from the ground-truth local signal PSD for continuous echo at low SER.

Figure 3 shows improvements of the WER index for almost all methods and configurations compared to the unprocessed microphone signal. As a comparison, with this ASR tool the clean local signals reaches average WER as low as 4.3% which is theoretically the best value to target with perfect echo cancellation. However, the simulated communication use-case appears more challenging here with globally higher WER. These reflect insertions in the the transcription of words from the reference signal and still recognized from the residual echo, explaining values above 100%

Figure 2 and fig. 3 show globally improved performance over the baseline [12]. Results with ground-truth local signal PSD also suggest that the proposed approach can still be improved with a better estimation of the local signal spectral content.

Figure 4 displays ERLE across time on a short excerpt for both use-cases (tested SER is here -20 dB). Bottom plot on fig. 4a shows that the proposed approach brings improvements over the baseline for both echo only and double-talk time periods. Figure 4b reports ERLE for the vocal assistant (continuous echo) with a sudden acoustic path change happening at t=20 s. This plot shows better performance (around 25 dB ERLE) compared to the baseline coming at a cost of slightly slower convergence after the acoustic path change.

V. Conclusion

We introduced a new approach combining a partition block frequency-domain adaptive filtering with a specific filter update normalization to achieve acoustic echo cancellation in double-talk situations. The proposed method, relying on instantaneous Signal-to-Echo Ratio estimation, brings substantial improvement compared to [12] without the need of double-talk detectors or prior information on the local signal. We demonstrated the algorithm usefulness when used in a signal-processing pipeline using speech recognition task. Experimental results also showed improved ERLE and speech intelligibility objective prediction which could probably still be improved within a full echo cancellation setup including residual echo suppressor. Future work could include perceptual assessment and more challenging situations [20] as well as testing such an approach within the Iterated Partitioned Block Frequency–Domain Adaptive Filtering framework [21].

VI. Acknowledgements

The authors thank Patrice Collen for precious discussions and advice on automated speech recognition tools and data.

References

[1] S. S. Haykin, B. Widrow, and B. Widrow, *Least-mean-square adaptive filters*. Wiley Online Library, 2003, vol. 31.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay *et al.*, "Advances in network and acoustic echo cancellation," 2001.

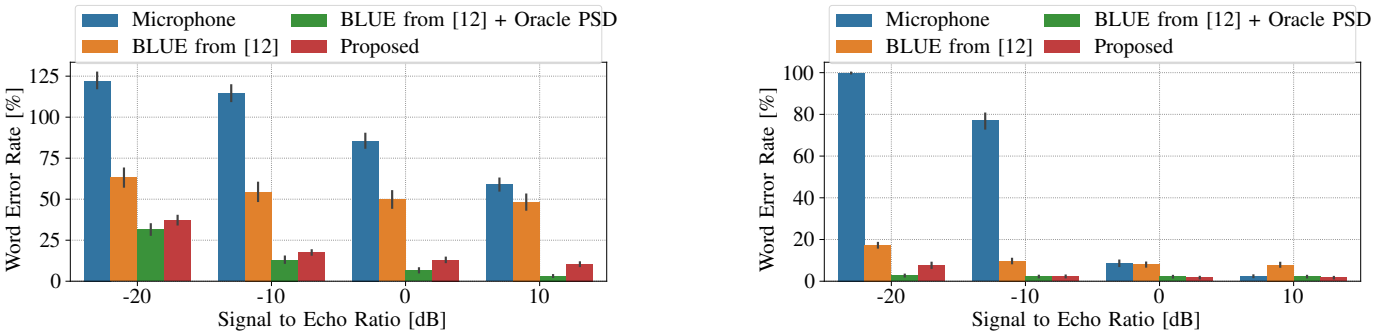
[3] H. K. Jung, N. S. Kim, and T. Kim, "A new double-talk detector using echo path estimation," *Speech communication*, vol. 45, no. 1, pp. 41–48, 2005.

[4] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1633–1644, 2006.

[5] M. A. Iqbal and S. L. Grant, "Novel variable step size nlms algorithms for echo cancellation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 241–244.

[6] T. van Waterschoot, G. Rombouts, P. Verhoeve, and M. Moonen, "Double-talk-robust prediction error identification algorithms for acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 846–858, 2007.

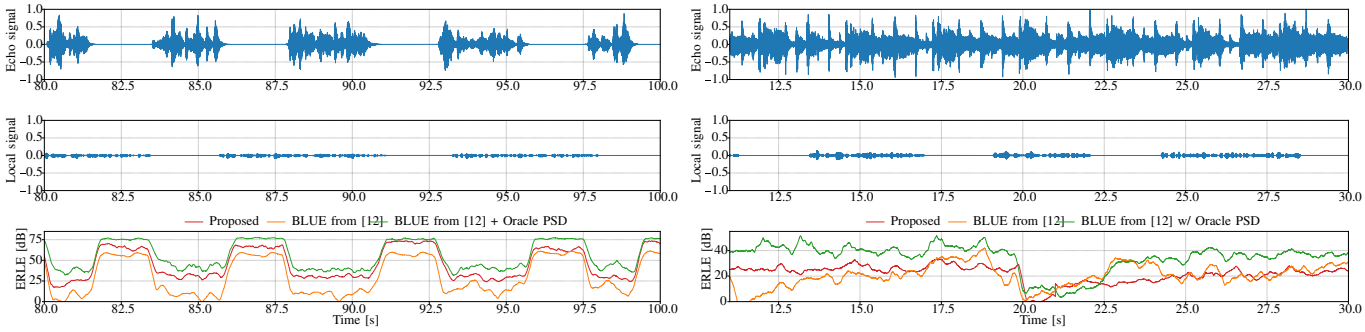
[7] J. M. Gil-Cacho, T. Van Waterschoot, M. Moonen, and S. H. Jensen, "A frequency-domain adaptive filter (fdaf) prediction error method (pem) framework for double-talk-robust acoustic echo cancellation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2074–2086, 2014.



(a) Communication

(b) Vocal assistant

Fig. 3: Signal to Echo Ratio *vs* Word Error Rate



(a) Communication

(b) Vocal assistant

Fig. 4: Time evolution of Echo Return Loss Enhancement

[8] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[9] J. P. Borrallo and M. G. Otero, "On the implementation of a partitioned block frequency domain adaptive filter (pbfdaf) for long acoustic echo cancellation," *Signal Processing*, vol. 27, no. 3, pp. 301–315, 1992.

[10] J. J. Shynk *et al.*, "Frequency-domain and multirate adaptive filtering," *IEEE Signal processing magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[11] T. Van Waterschoot and M. Moonen, "Double-talk robust acoustic echo cancellation with continuous near-end activity," in *2005 13th European Signal Processing Conference*. IEEE, 2005, pp. 1–4.

[12] T. Trump, "A frequency domain adaptive algorithm for colored measurement noise environment," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 3. IEEE, 1998, pp. 1705–1708.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[14] L. F. Larnel, J.-L. Gauvain, and M. Eskenazi, "Bref, a large vocabulary spoken corpus for french," in *Second european conference on speech communication and technology*, 1991.

[15] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical and jazz music databases." in *Ismir*, vol. 2, 2002, pp. 287–288.

[16] J. Y. Wen, N. D. Gaubitch, E. A. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the mardy database," in *in Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC*. Citeseer, 2006.

[17] "Google WebRTC," Accessed on 12/. [Online]. Available: https://webrtc.org/

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[19] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[20] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets and testing framework," *arXiv preprint arXiv:2009.04972*, 2020.

[21] K. Eneman and M. Moonen, "Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 143–158, 2003.