AN ACOUSTIC KEYSTROKE TRANSIENT CANCELER FOR SPEECH COMMUNICATION TERMINALS USING A SEMI-BLIND ADAPTIVE FILTER MODEL

Herbert Buchner[1]*, Jan Skoglund*[2]*, and Simon Godsill*[1]

[1] University of Cambridge, Information Engineering Division, CB2 1PZ Cambridge, UK
[2] Google Inc., Mountain View, CA 94043, USA

ABSTRACT

In many teleconferencing applications using modern laptop and netbook devices it is common to encounter annoying keyboard typing noise. In this paper we propose an acoustic keystroke transient canceler for speech communication terminals as a novel broadband adaptive filter application in such a hands-free scenario. We present this approach in the context of the Google Chromebook Pixel device which is equipped with a special audio reference channel providing various new signal processing possibilities. Our novel semi-blind/semi-supervised approach exploiting this new degree of freedom, combined with the system-based broadband estimation and a novel adaptation control yields a high-quality speech enhancement even under challenging acoustic conditions.

Index Terms— Acoustic keystroke transients, noise reduction, impulsive noise, hands-free speech communication.

1. INTRODUCTION

The rapid increase in availability of high speed internet connections has made personal computing devices a very popular basis for teleconferencing applications. While the embedded microphones, loudspeakers and webcams in laptop or tablet computers make setting up conference calls very easy, the resulting acoustic hands-free communication scenario generally brings with it the need for a number of challenging signal processing problems, such as acoustic echo control, signal separation/extraction from background noise or other competing sources, and, ideally, dereverberation [1, 2].

A special type of acoustic noise which can be a particularly persistent problem, and which we deal with in the present paper, is the impulsive noise caused by keystroke transients, especially when using the embedded keyboard of a laptop computer during teleconferencing applications (e.g., in order to make notes or to write emails). In such a setup, this impulsive noise in the microphone signals can be a significant nuisance due to the spatial proximity between the microphones and the keyboard, and partly due to possible vibration effects and solid-borne sound conduction within the device casing.

Most of the well-known single-channel speech enhancement algorithms are typically based on noise power estimation and spectral amplitude modification in the short-time Fourier transform (STFT) domain, e.g., [3]. However, reducing highly nonstationary noise such as keystroke transients remains a challenging problem for many algorithms of this type and a still very active field of research, e.g., [4],[5]. In a recent study, the application of separation methods like non-negative matrix factorization (NMF) in the spectral domain has shown promising results for impulsive noise [6]. While this can be effective where long signal samples are available, particularly for batch estimation, unfortunately, in practice there is very little adaptation time available due to the short activity of the key stroke transients and the variations of the acoustic click events. Note also that the keyboard noise is broadband with its dominant frequency components typically in the same range as that of the speech signal. Due to these challenging conditions, this signal processing problem has been tackled so far mainly by missing feature approaches, e.g., [7] based on [8]. Similar approaches are also known from image and video processing, e.g. [9]. Similarly to the speech enhancement algorithms mentioned above, the missing feature-type algorithms typically require very accurate detections of the keystroke transients. Moreover, in the case of keystroke noise, this detection problem is exacerbated by both the reverberation effects and the fact that each keystroke actually leads to two audible clicks with unknown and varying distance, whereby the peak of the second click is often buried entirely in the overlapping speech signal. Note that simply using the typing information from the operating system of the device is usually not accurate enough as the temporal deviation between the typing information registered by the OS and the actual acoustic event can vary widely and is not deterministic. Recent papers addressing the detection problem for suppression of transients (but otherwise still based on a conventional (signal-based) noise suppression algorithm or a modification thereof) are, e.g., [10, 11, 12, 13].

The purpose of this paper is twofold. First, we will clarify and analyze the signal processing problem in somewhat more detail and we will focus on the specific class of approaches characterised by the use of *broadband adaptive FIR filters*. In this context we specifically turn our attention to the Chromebook Pixel which is one of the first commercial of-the-shelf products featuring an additional reference sensor underneath the keyboard. As we will see, in this context, the related semi-supervised / semi-blind signal processing problem can be regarded as a new class of adaptive filtering problems in the hands-free context in addition to the already more extensively studied classes of problems in this field [1, 2]. Since the operating system of the Chromebook is based on the well-documented open-source project Chromium OS [14], it provides new opportunities for the development of further improved signal enhancement algorithms. Secondly, based on this available platform, we discuss and evaluate in this paper a novel candidate algorithm specifically for semi-supervised acoustic keystroke transient cancellation.

2. KEYBOARD REFERENCE PICKUP

Signal reconstruction generally becomes more and more challenging with increasing typing speed and/or increasing room reverberation causing the effects of the keystrokes to overlap. In reverberant environments, the click noise may well extend over multiple analysis blocks. Note that each character consists of two keystroke transients (downward and upward movements of the keys.) A statistical analysis of typical typing speeds can be found, e.g., in [15].

Two ways to tackle the above problems are (1) to take into account some less defective signal as side information on the

keystrokes and (2) to take into account the acoustic signal propagation including the reverberation effects using dynamical models. Hence, the novel approach in this paper takes advantage of an additional microphone underneath the keyboard and, as shown in the next sections, it uses an adaptive filtering approach exploiting the knowledge of this keyboard microphone signal. The Google Chromebook Pixel exhibits such a reference microphone as well as two voice microphones on top of the display, as shown in Fig. 1. This setup allows for the development of more powerful, semi-supervised algorithms.

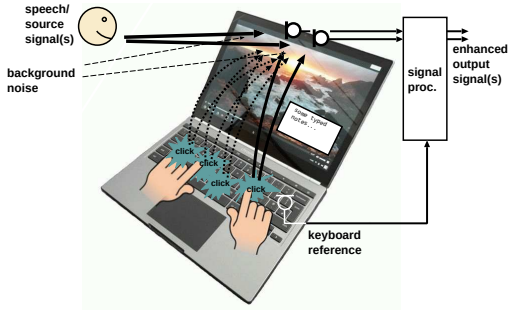

Fig. 1. Keyboard reference pickup in Google's chromebook pixel.

With the available reference signal and the application of adaptive filtering, the problem appears to be similar to a conventional acoustic echo cancellation (AEC) problem [16, 17, 18] or an interference cancellation problem [19]. However, there are notable differences between conventional AEC and this keystroke transient suppression which are reflected by the following *requirements*:

(i) The "echo path" to be identified is rapidly time varying.

(ii) The excitation (keystroke transients) of the "echo path" is typically very short, i.e., limited data for the estimation process.

(iii) We have cross-talk of low (but noticeable) power from the speech source into the keybed microphone.

(iv) Double-talk control (or double-talk detection in particular), as in conventional AEC is not straightforward here (mainly due to (iii) and (v)).

(v) Highly nonlinear systems. Our experiments have shown that the acoustic paths from the keyboard to the microphones contain significant nonlinear contributions due to the solid-borne sound conduction within the casing. The nonlinear contributions (rattling) also exhibit a significant memory [20].

(vi) The algorithm should have low complexity despite the challenging requirements (i)-(v).

In the next Sections, we develop and evaluate an efficient semi-supervised algorithm which is designed to meet the above requirements (i)-(vi) for keystroke transient cancellation.

3. KEYSTROKE TRANSIENT CANCELLATION BASED ON BROADBAND ADAPTIVE MIMO FILTERING

Our setup can be regarded as an acoustic mixing system with multiple input channels and multiple output channels (MIMO) consisting of impulse responses h_{qp}, $q = 1, 2$, $p = 1, 2, 3$, see Fig. 2. Analogously, the signal processing for extracting the desired speech signal can be considered as a corresponding MIMO demixing system. The coefficients of the MIMO system are then regarded as latent variables which are assumed to have less variability over multiple time frames of the observed data. As they allow for a global optimization over longer data sequences, latent variable models have the well-known advantage of reducing the dimensions of data, making it easier to understand and, thus, in our application, reduce or avoid distortions in the output signals. In the reminder of this paper we will refer to this approach as *system-based* optimization in contrast to the existing *signal-based* approaches mentioned in Sect. 1. Note that in practice we can expect synergies by combining system-based and signal-based approaches for signal enhancement as in AEC [18].

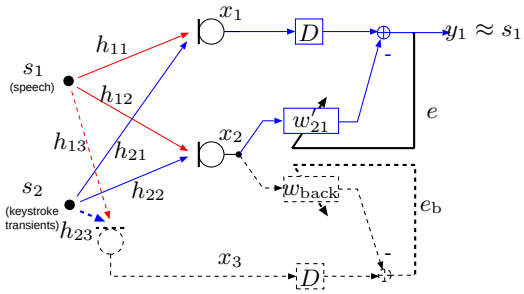

Fig. 2. Block diagram of the proposed simplified system using adaptive filters in the foreground and in the background for adaptation control (according to Sect. 4).

Based on the framework of broadband adaptive MIMO signal processing there are several conceivable *specialized* adaptive filter configurations to efficiently solve the problem. These specialized approaches, such as the already mentioned AEC approach can often be obtained as 'pruned' versions of the full blind source separation (BSS) approach, e.g., [21, 22, 23, 24]. In this paper, we focus on one particularly simple approach that falls into this category but due to the above list of requirements (i)-(vi), it differs from the AEC setup in order to largely circumvent the nonlinearity issues.

We can observe that the relation between x_1, x_2 is closer to linearity than the relation between x_3, x_1 and the relation between x_3, x_2, respectively. This would motivate a blind spatial signal processing using the two array microphones x_1, x_2.

On the other hand, x_3 still contains significantly less crosstalk and less reverberation due to the proximity between the keyboard and the keyboard microphone. Therefore, the keyboard microphone is best suited for guiding the adaptation. In other words, while the core algorithm is adapted blindly, the overall system can be considered as a *semi-blind* system. The guidance of the adaptation using the keyboard microphone addresses both the double-talk problem and the resolution of the inherent permutation ambiguity concerning the desired source in the output of blind adaptive filtering algorithms.

Interestingly, the resulting structure can be interpreted either as a subspace approach/blind signal extraction (BSE) approach [25] or as a method for blind system identification (BSI) for single-input and multiple-output (SIMO) systems [23, 24] (In the second-order-statistics case this pruned case coincides rigorously with the blind SIMO system identification approach in, e.g., [26, 27]). As we will see, both interpretations are utilized in our practical implementation of the overall system; the BSE for extracting the desired speech signal, and the BSI for the novel double-talk control proposed in Sect. 4.

In this paper, we further simplify the demixing system by keeping the subfilter w_{11} fixed as a pure delay, as shown in Fig. 2. In this way we avoid any further *linear* distortions of the desired speech signal. The adaptive subfilter w_{21} is then simply modified to w'_{21}. Note that this approach can be considered as the above SIMO-based BSE approach with an additional equalizer. To ensure causality of the adaptive filter w'_{21} for arbitary speaker positions, the delay is chosen as $D = \lceil L/2 \rceil$, where L denotes the demixing filter length.

4. AN EFFICIENT REALIZATION AND CONTROL OF THE ADAPTATION

Having identified a promising candidate for an optimal system-based approach according to the above requirements (i)-(vi), we are now going to develop the ingredients for an efficient practical realization. Special emphasis will be placed on the novel semi-blind double-talk detection specifically taking into account requirements (iii)-(v). Later, in Table 1, we will summarize a pseudocode based on these building blocks and the semi-blind structure according to Fig. 2.

4.1. Broadband block-online frequency-domain adaptation

As a first step, we apply a computationally efficient frequency-domain formulation of the above filter structure following [28]. An important feature of this frequency-domain framework is that it increases the efficiency of both the adaptatation processes (approximate diagonalization of the Hessian) and the filtering process (fast convolution by exploiting the efficiency of the FFT) [28].

In this paper we will work with *partitioned blocks*, i.e., the (integer) block length $N = L/K$ can be a fraction of the filter length L. This decoupling of L and N is especially desirable for handling highly non-stationary signals such as the keystroke transients in our application [29, 28, 30].

Let us consider the input-output relationship for one individual demixing subfilter w_{pq}. The output signal of this subfilter at time n reads $y_{qp}(n) = \sum_{\ell=0}^{L-1} x_p(n-\ell) w_{pq,\ell}$, where $w_{pq,\ell}$ are the coefficients of the filter impulse response w_{pq}. By partitioning the impulse response w_{pq} of length L into K segments of integer length $N = L/K$ as in [29], $y_{qp}(n)$ can be written as

$$y_{qp}(n) = \sum_{k=0}^{K-1} \mathbf{x}_{p,k}^{\mathrm{T}}(n) \mathbf{w}_{pq,k}, \qquad (1)$$

where vector $\mathbf{x}_{p,k}(n)$ consists of $x_p(n-Nk)$, ..., $x_p(n-Nk-N+1)$. Superscript $^{\mathrm{T}}$ denotes transposition of a vector or a matrix. The length-N vectors $\mathbf{w}_{pq,k}$, $k = 0, \ldots, K-1$ represent subfilters of the partitioned tap-weight vector $\mathbf{w}_{pq} = [\mathbf{w}_{pq,0}^{\mathrm{T}}, \ldots, \mathbf{w}_{pq,K-1}^{\mathrm{T}}]^{\mathrm{T}}$.

Since our algorithm is based on block processing, we consider blocks of output samples. A length-N block of output samples

$$\mathbf{y}_{qp}(m) = [y_{qp}(mN), \ldots, y_{qp}(mN+N-1)]^{\mathrm{T}} \qquad (2)$$

can, analogously to (1), be formulated as a sum of K matrix-vector products (instead of the inner products in (1)). A key insight in broadband frequency-domain adaptive filtering is that each of the K matrices can be *diagonalized* very efficiently and without any approximation using DFT matrices (implemented by FFTs) after applying certain window matrices, as shown in detail, e.g., in [28]. These window matrices and the resulting closed-form expressions (such as, e.g., the block error signal $\mathbf{e}(m)$, with m being the block time index) are summarized in Table 1.

Having expressed the error signal in a compact partitioned-block frequency-domain notation, similarly compact formulations of frequency-domain adaptive filter (FDAF) algorithms, based on a block-based optimization criterion, such as $J(m, \underline{\mathbf{w}}) = (1-\lambda) \sum_{i=0}^{m} \lambda^{m-i} \mathbf{e}^{\mathrm{T}}(m) \mathbf{e}(m)$, where λ is a forgetting factor $0 < \lambda < 1$, can be derived [28]. By exploiting the efficiency of the FFT, the resulting FDAF algorithms exhibit both a very low complexity ($O(\log L)$ per sample), and a fast RLS-like convergence speed as necessary in our application.

Table 1 shows the pseudocode of an overall algorithm based on the system configuration according to Fig. 2, and the multidelay formulation mentioned above. As indicated in Fig. 2, the overall system consists of a *foreground filter* (i.e., the main adaptive filter producing the enhanced output signal y_1 as described above) and a separate *background filter* (dashed part, used for controlling the adaptation of the foreground filter). These two components are also reflected by the two lowermost (main) sections in the pseudocode. The foreground filter corresponds to the steps (3s)-(3w), i.e., the last section in the pseudocode, including the necessary Kalman gain (3e),(3f) [which is used for computational efficiency for both the foreground filter and background filter due to their common input signal $\mathbf{X}_2(m)$], and the required input signals (3a)-(3c). The background filter for adaptation control will be discussed in more detail below in Sect. 4.2.

An important feature of the implementation according to Table 1 in order to further speed up the convergence are the additional offline iterations (index ℓ) in each block. This method also follows from the relation between supervised and blind adaptive filtering [21], where it is probably more common. Moreover, to avoid the undesirable 'overlearning' phenomenon (from the viewpoint of system-based estimation) for a high number of offline iterations with this method, yet allow to a certain degree for the exploitation of its rapid tracking capability of local signal statistics, the total number $\ell_{\max}$ of offline iterations is subdivided into two steps.

4.2. Semi-blind multidelay double-talk detection

We now focus on the important aspect of controlling the adaptation, i.e., the double-talk detector (first main part in Table 1). Here, the goal is to develop a reliable decision mechanism so that the adaptation of the keystroke transient canceller is performed *only during the exclusive* activity of the keystroke transients.

Despite of the availability of the keyboard reference microphone (signal x_3), it turns out that in our scenario a reliable adaptation control is a more challenging task than the adaptation control problem for the well-known supervised adaptive filtering case, e.g., for acoustic echo cancellation. This is mainly due to the cross-talk of the desired speech signal into the keyboard reference microphone, as well as the very significant nonlinear components in the propagation paths of the keystroke transients (requirements (iii)-(v) in Sect. 2).

Hence, in addition to a power-based or correlation-based decision statistic (as for AEC; see, e.g., [30] and references therein), we propose in Table 1 a novel adaptation control based on multiple decision criteria which also exploit the spatial selectivity by the multiple microphone channels. Indeed, the resulting algorithm can be regarded as an according semi-blind generalization of the multidelay-based detection mechanism proposed earlier in [30].

In addition to the short-time signal power $\sigma_{x_3}^2(m)$ as a first detection variable, the detection variable ξ_1 describes the ratio of a linear approximation to the nonlinear contribution in x_3. Note that this detection statistic formally resembles the one in [30] but the mechanism here is slightly different.

Probably the most important criterion is described by the detection variable ξ_2. This criterion can be seen as a spatio-temporal source signal activity detector. Note that both, the detection variables ξ_1 and ξ_2 are based on the adaptive background filter (similar to the foreground filter, but with slightly larger stepsize and smaller forgetting factor for quick reaction of the detection mechanism).

The detection variable ξ_2 exploits the microphone array geometry and the fact that the background filter performs an *approx.* BSI of h_{13} and/or h_{23} (see Sect. 3). According to the physical setup, we can safely assume that the direct path of h_{23} is always significantly shorter than the direct path of h_{13}. Due to the relation of the maxima of the background filter coefficients and the time difference of arrival (e.g., [23]), an approximate decision on the activity of both sources s_1 and s_2 can be made ($1 \leq a < b < c \leq L$ in (3p)). To further improve the detection accuracy, a regularization for sparse

learning of the background filter coefficients is applied ((3m)-(3o), where $\Phi(\cdot, a)$ denotes a center clipper, also called *shrinkage operator*, of width a) [31, 32].

5. EXPERIMENTAL EVALUATION

The signal extraction has been evaluated using recorded signals from the Chromebook in a regular office room ($T_{60} \approx 300$ms) by both objective and subjective performance measures. Figure 3 illustrates a typical use case. The first subplot shows one of the recorded voice mic signals. During the first 6sec. only the speech signal was active (male speaker), and thereafter the typing activity sets in (various keys across the whole keyboard were used). The simulation covers single-talk and double-talk (6..10sec and 14..16.5sec) conditions. The second and third subplots show the processed output signal and the acoustic click suppression (dB), respectively. We see that both, with and without simultaneous speech activity, a significant suppression of the click noise can be achieved and maintained.

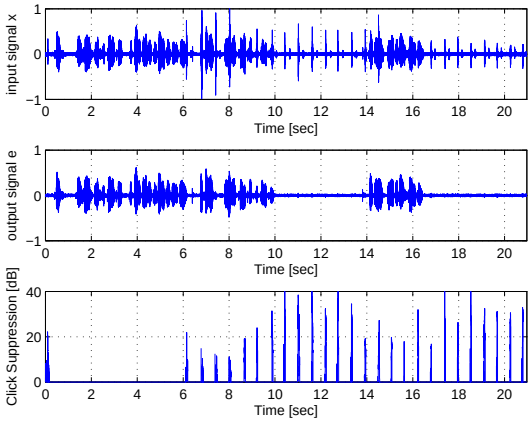

Fig. 3. Signal extraction result.

Figure 4 shows a summary of the results of a standardized subjective listening test (MUSHRA, 'MUlti Stimulus test with Hidden Reference and Anchor' [33]) carried out with 10 listeners. The sound quality was quantified on a scale from 0 to 100 for 4 different use cases with two different speech signals and two different typing speeds. They all contained single-talk and double-talk situations as illustrated above.

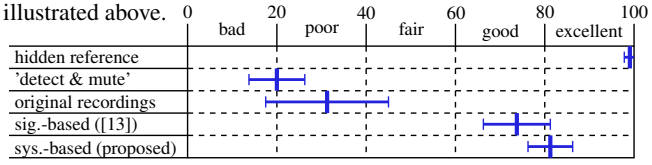

Fig. 4. Results of the MUSHRA listening test (average and 95% confidence intervals).

Besides taking into account various use cases in this evaluation, our main objective here was to compare a recently presented state-of-the-art *signal-based approach* [13] with the novel *system-based approach* presented in this paper. It is visible from the graph that the system-based approach is able to outperform the quality of the previous algorithm due to the built-in system model. Moreover, due to the different estimation mechanisms of the two approaches, a synergistic combinability for even further improvement is to be expected.

6. CONCLUSIONS

In this paper we have presented a novel and efficient semi-blind system-based approach for keystroke transient cancellation. The approach is highly efficient, yields a high suppression performance, and minimal signal distortion. A possible extension to the approach is, e.g., the use of robust statistics [30, 34, 35].

Table 1: Robust semi-blind EMDF and semi-blind EMD DTD

Definition of window matrices:

$$\begin{aligned}
\mathbf{W}^{01}_{N\times 2N} &= [\ \mathbf{0}_{N\times N}\quad \mathbf{I}_{N\times N}\ ]\\
\mathbf{W}^{10}_{2N\times N} &= [\ \mathbf{I}_{N\times N}\quad \mathbf{0}_{N\times N}\ ]^{\mathrm{T}}\\
\mathbf{W}^{01}_{2N\times 2N} &= \begin{bmatrix} \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}\\
\mathbf{G}^{01}_{2N\times 2N} &= \mathbf{F}_{2N}\mathbf{W}^{01}_{2N\times 2N}\mathbf{F}^{-1}_{2N}\\
\mathbf{W}^{10}_{2N\times 2N} &= \begin{bmatrix} \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\end{bmatrix}\\
\tilde{\mathbf{G}}^{10}_{2N\times 2N} &= \mathbf{F}_{2N}\mathbf{W}^{10}_{2N\times 2N}\mathbf{F}^{-1}_{2N}\\
\mathbf{G}^{10}_{2L\times 2L} &= \mathrm{diag}\{\tilde{\mathbf{G}}^{10}_{2N\times 2N},\ldots,\tilde{\mathbf{G}}^{10}_{2N\times 2N}\}
\end{aligned}$$

Input signals:

$$\begin{aligned}
\mathbf{x}_1(m) &= [x_1(mN-D),\ldots\\
&\quad \ldots, x_1(mN-D+N-1)]^{\mathrm{T}} && (3a)\\
\mathbf{X}_{2,k}(m) &= \mathrm{diag}\{\mathbf{F}_{2N}[x_2(mN-Ni-N),\ldots\\
&\quad \ldots, x_2(mN-Ni+N-1)]^{\mathrm{T}}\},\\
&\quad k=0,\ldots,K-1 && (3b)\\
\mathbf{X}_2(m) &= [\mathbf{X}_{2,0}(m),\mathbf{X}_{2,1}(m),\ldots,\mathbf{X}_{2,K-1}(m)] && (3c)\\
\underline{\mathbf{x}}_3(m) &= \mathbf{F}_{2N}[\mathbf{0}_{1\times N}, x_3(mN-D),\ldots\\
&\quad \ldots, x_3(mN-D+N-1)]^{\mathrm{T}} && (3d)
\end{aligned}$$

Kalman gain:

$$\begin{aligned}
\mathbf{S}'(m) &= \lambda\mathbf{S}'(m-1)+(1-\lambda)\mathbf{X}_2^{\mathrm{H}}(m)\mathbf{X}_2(m) && (3e)\\
\mathbf{K}(m) &= \mathbf{S}'^{-1}(m)\mathbf{X}_2{}^{\mathrm{H}}(m) && (3f)
\end{aligned}$$

Double-talk detector (background filter):

$$\begin{aligned}
\underline{\mathbf{w}}^0_{\mathrm{b}}(m) &:= \underline{\mathbf{w}}_{\mathrm{b}}(m-1)\\
&\textbf{for } \ell=1,\ldots,\ell_{\max,\mathrm{sys,back}}:\\
\underline{\mathbf{e}}^{\ell}_{\mathrm{b}}(m) &= \underline{\mathbf{x}}_3(m)-\mathbf{G}^{01}_{2N\times 2N}\mathbf{X}_2(m)\hat{\underline{\mathbf{w}}}^{\ell-1}_{\mathrm{b}}(m) && (3g)\\
\underline{\mathbf{w}}^{\ell}_{\mathrm{b}}(m) &= \underline{\mathbf{w}}^{\ell-1}_{\mathrm{b}}(m)+\\
&\quad +\mu_{\mathrm{b}}2(1-\lambda_{\mathrm{b}})\mathbf{G}^{10}_{2L\times 2L}\mathbf{K}(m)\underline{\mathbf{e}}^{\ell}_{\mathrm{b}}(m) && (3h)\\
&\textbf{end for}\\
\underline{\mathbf{w}}'_{\mathrm{b}}(m) &:= \underline{\mathbf{w}}^{\ell_{\max,\mathrm{sys,back}}}_{\mathrm{b}}(m)\\
\sigma^2_{x_3}(m) &= \lambda_{\mathrm{b}}\sigma^2_{x_3}(m-1)+(1-\lambda_{\mathrm{b}})\underline{\mathbf{x}}_3^{\mathrm{H}}(m)\underline{\mathbf{x}}_3(m) && (3i)\\
\mathbf{s}_k(m) &= \lambda_{\mathrm{b}}\mathbf{s}_k(m-1)+(1-\lambda_{\mathrm{b}})\mathbf{X}^*_{2,k}(m)\underline{\mathbf{x}}_3(m),\\
&\quad k=0,\ldots,K-1 && (3j)\\
\xi_1(m) &= \frac{\sum_{k=0}^{K-1}\underline{\mathbf{w}}'_{\mathrm{b},k}{}^{\mathrm{H}}(m)\mathbf{s}_k(m)}{\sigma^2_{x_3}(m)} && (3k)\\
\mathbf{w}'_{\mathrm{b}}(m) &= \mathrm{diag}\{\mathbf{W}^{01}_{N\times 2N}\mathbf{F}^{-1}_{2N},\ldots,\mathbf{W}^{01}_{N\times 2N}\mathbf{F}^{-1}_{2N}\}\times\\
&\quad \times\underline{\mathbf{w}}'_{\mathrm{b}}(m) && (3l)\\
\mathbf{w}_{\mathrm{b}}(m) &= (1-2\lambda_{\mathrm{r}}\mu_{\mathrm{b}})\mathbf{w}'_{\mathrm{b}}(m)-\\
&\quad -2\lambda_{\mathrm{r}}\mu_{\mathrm{b}}\left(\mathbf{b}_{\mathrm{r}}(m-1)-\mathbf{d}_{\mathrm{r}}(m-1)\right) && (3m)\\
[\mathbf{d}_{\mathrm{r}}(m)]_n &= \Phi\left([\mathbf{w}_{\mathrm{b}}(m)+\mathbf{b}_{\mathrm{r}}(m-1)]_n, \tfrac{\rho_{\mathrm{r}}}{2\lambda_{\mathrm{r}}}\right),\\
&\quad n=1,\ldots,N && (3n)\\
\mathbf{b}_{\mathrm{r}}(m) &= \mathbf{b}_{\mathrm{r}}(m-1)+\mathbf{w}_{\mathrm{b}}(m)-\mathbf{d}_{\mathrm{r}}(m) && (3o)\\
\xi_2(m) &= \frac{\max_{a\le i\le b}|w_{\mathrm{b},i}(m)|}{\max_{b<i\le c}|w_{\mathrm{b},i}(m)|} && (3p)\\
\underline{\mathbf{w}}_{\mathrm{b}}(m) &= \mathrm{diag}\{\mathbf{F}_{2N}\mathbf{W}^{10}_{2N\times N},\ldots,\mathbf{F}_{2N}\mathbf{W}^{10}_{2N\times N}\}\times\\
&\quad \times\mathbf{w}_{\mathrm{b}}(m) && (3q)\\
&\textbf{if } \xi_1\ge T_1 \ \&\ \xi_2<T_2 \ \&\ \sigma^2_{x_3}(m)>T_3\\
\mu' &= \mu(1-\lambda) \quad \text{('single-talk' } \Rightarrow \text{ adapt foreground)} && (3r)\\
&\textbf{else}\\
\mu' &= 0 \quad \text{('double-talk' } \Rightarrow \text{ don't adapt foregr.)}\\
&\textbf{end if}
\end{aligned}$$

Keystroke transient canceller (foreground filter):

$$\begin{aligned}
\underline{\mathbf{w}}^0(m) &:= \underline{\mathbf{w}}(m-1)\\
&\textbf{for } \ell=1,\ldots,\ell_{\max,\mathrm{sys}}:\\
\mathbf{e}^{\ell}(m) &= \mathbf{x}_1(m)-\mathbf{W}^{01}_{N\times 2N}\mathbf{F}^{-1}_{2N}\mathbf{X}_2(m)\underline{\mathbf{w}}^{\ell-1}(m) && (3s)\\
\underline{\mathbf{w}}^{\ell}(m) &= \underline{\mathbf{w}}^{\ell-1}(m)+\mu'\mathbf{G}^{10}_{2L\times 2L}\mathbf{K}(m)\times\\
&\quad \times\mathbf{F}_{2N}\mathbf{W}^{01}_{2N\times N}\mathbf{e}^{\ell}(m) && (3t)\\
&\textbf{end for}\\
\underline{\mathbf{w}}(m) &:= \underline{\mathbf{w}}^{\ell_{\max,\mathrm{sys}}}(m) && (3u)\\
&\textbf{for } \ell=\ell_{\max,\mathrm{sys}}+1,\ldots,\ell_{\max}:\\
\mathbf{e}^{\ell}(m) &= \mathbf{x}_1(m)-\mathbf{W}^{01}_{N\times 2N}\mathbf{F}^{-1}_{2N}\mathbf{X}_2(m)\underline{\mathbf{w}}^{\ell-1}(m) && (3v)\\
\underline{\mathbf{w}}^{\ell}(m) &= \underline{\mathbf{w}}^{\ell-1}(m)+\mu'\mathbf{K}(m)\mathbf{F}_{2N}\mathbf{W}^{01}_{2N\times N}\mathbf{e}^{\ell}(m) && (3w)\\
&\textbf{end for}\\
\mathbf{y}_1(m) &:= \mathbf{e}^{\ell_{\max}}(m)
\end{aligned}$$

7. REFERENCES

[1] W. Kellermann, H. Buchner, W. Herbordt, and R. Aichner, "Multichannel acoustic signal processing for human/machine interfaces - fundamental problems and recent advances," in *Conf. Rec. 18th Int. Congress on Acoustics*, Kyoto, Japan, Apr. 2004.

[2] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin: Springer, 2006.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[4] J. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1112–1123, Aug. 2008.

[5] S. Godsill, "The shifted inverse-gamma model for noise-floor estimation in archived audio recordings," *Signal Processing*, vol. 90, pp. 991–999, 2010.

[6] N. Mohammadiha and S. Doclo, "Transient noise reduction using nonnegative matrix factorization," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014.

[7] A. Subramanya, M. Seltzer, and A. Acero, "Automatic removal of typed keystrokes from speech signals," *IEEE SP Letters*, vol. 14, no. 5, pp. 363–366, May 2007.

[8] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[9] K. Meisinger and A. Kaup, "Spatiotemporal selective extrapolation for 3-D signals and its applications in video communications," *IEEE Trans. on Image Processing*, vol. 16, no. 9, Sept. 2007.

[10] A. Abramson and I. Cohen, "Enhancement of speech signals under multiple hypotheses using an indicator for transient noise presence," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2007, pp. 553–556.

[11] A. Sugiyama, "Single-channel impact-noise suppression with no auxiliary information for its detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2007.

[12] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[13] S. Godsill, H. Buchner, and J. Skoglund, "Detection and suppression of keyboard transient noise in audio streams with auxiliary keybed microphone," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015.

[14] The Chromium projects website, "Chromebook pixel," http://www.chromium.org/chromium-os/developer-information-for-chrome-os-devices/chromebook-pixel.

[15] T. Ostrach, "Typing speed: How fast is average," Orlando, FL, USA, 1997, http://orbitouch.com/wp-content/uploads/2012/03/Average-OrbiTouch-Typing-Speed.pdf
or http://readi.info/documents/TypingSpeed.pdf.

[16] C. Breining, P. Dreiseitel, E. Hänsler, 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*, pp. 42–69, Jul. 1999.

[17] J. Benesty, T. Gänsler, D. Morgan, M. Sondhi, and S. Gay, *Advances in Network and Acoustic Echo Cancellation*. Berlin: Springer, 2001.

[18] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic Press Library in Signal Processing*, R. Chellappa and S. Theodoridis, Eds. Elsevier/Academic Press, 2014.

[19] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NJ: Prentice Hall Inc., 2002.

[20] A. Birkett and R. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 1995.

[21] H. Buchner and W. Kellermann, "A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, May 2008.

[22] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Montreal, Canada, May 2004, pp. 889–892.

[23] ——, "TRINICON-based blind system identification with application to multiple-source localization and separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and S. Sawada, Eds. Berlin: Springer, Sept. 2007, pp. 101–147.

[24] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. Naylor and N. Gaubitch, Eds. London: Springer, Jul. 2010, pp. 311–385.

[25] H. Buchner and K. Helwani, "On the relation between blind system identification and subspace tracking and associated generalizations," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Nov. 2010.

[26] M. Gürelli and C. Nikias, "EVAM: an eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 134–149, Jan. 1995.

[27] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, pp. 384–391, Jan. 2000.

[28] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation," in *Adaptive signal processing: Application to real-world problems*, J. Benesty and Y. Huang, Eds. Berlin: Springer, Jan. 2003, pp. 95–128.

[29] J.-S. Soo and K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 373–376, Feb. 1990.

[30] H. Buchner, J. Benesty, T. Gänsler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 9, Sept. 2006.

[31] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[32] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing," *SIAM J. Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[33] ITU-R, "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," Jan. 2003.

[34] P. Huber, *Robust Statistics*. New York: Wiley, 1981.

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