BAYESIAN INFERENCE MODEL FOR APPLICATIONS OF TIME-VARYING ACOUSTIC SYSTEM IDENTIFICATION

Gerald Enzner

Institute of Communication Acoustics, Ruhr-University Bochum
D–44780 Bochum, Germany
Email: gerald.enzner@rub.de, Phone: +49-234-32-25392

ABSTRACT

A major challenge in acoustic signal processing lies in the uncertainty regarding the current state of the acoustic environment. The relevant applications in the field of speech and audio signal processing include the multichannel sound capture, the signal processing for spatial sound control, and the acoustic echo/interference cancellation. In this paper, a Bayesian impulse response model is proposed for acoustic system identification. It is justified by the stochastic nature of time-varying and noisy environments. In particular, we argue for a state-space dynamical model of the unknown impulse responses as a suitable form to incorporate *a priori* information of the acoustic environment. For the echo/interference cancellation case, we then describe the Bayesian inference of the acoustic system. It is structurally and experimentally compared to maximum-likelihood and least-squares estimators which are both rooted in deterministic system modeling. Algorithmic structure and performance, both speak for the Bayesian inference.

1. APPLICATIONS OF ACOUSTIC SYSTEM IDENTIFICATION

Figure 1 depicts a generic acoustic environment with audio reproduction and recording capabilities. The loudspeaker arrangement on the reproduction side presents acoustic scenes to the users of the environment as shown, e.g., by the virtual sound source in front of the loudspeakers. The loudspeaker driving signals can be derived from the received signals from a remote acoustic environment or provided through broadcast and storage media.

The users of the acoustic environment act in a two-fold way. On the one hand, they act as acoustic receivers, i.e., as consumers of the acoustic scene, indicated by the in-ear microphones in Fig. 1. On the other hand, the users naturally act as inner sources of the acoustic system, e.g., by their speech utterances to the recording microphones or simply by their internal voice communication.

On the recording side, Fig. 1 depicts several microphone channels. This multi-microphone arrangement can be used to capture the voices of users '1' and '2', or possibly more internal sources, e.g., to transmit to a remote environment or to feed a speech recognizer in the recording unit. The recording microphones might also include a set of in-ear transducers, as shown for user '1', in order to supply a perceptual reference of the soundfield.

Various use cases obviously arise from the capabilities of the acoustic environment and they might even take place simultaneously. In the following, we consider the related signal processing applications on the recording and reproduction side of the system and we address the undesired interference of both sides.

1.1 Multichannel Speech Acquisition

Among the recording side applications in acoustics, we have the dereverberation, denoising, localization, and separation of distant speech using microphone arrays. Comprehensive presentations of this field can be found, e.g., in [1, 2]. The solutions to these problems inherently require an exact or approximate equalization of the acoustic system between users '1' and '2' and the microphone transducers in Fig 1. This acoustic system can be represented, e.g., by a set of acoustic impulse responses which fully describes the undesired linear distortion and additive mixture of the original speech sources. System identification can give access to the impulse responses and with their availability we can apply the sophisticated designs of multi-channel speech enhancement algorithms to restore the original speech sources from the distorted microphone signals. The principle feasibility of multichannel acoustic equalization on the basis of the known system was demonstrated in [3].

Typically, the acoustic system is time-varying and we do not have a reference to the original source signals and therefore no means to perform supervised system identification [4]. In this case, we face the very difficult blind system identification problem, the solution of which can suffer from inherent source-filter ambiguities [5]. Blind system identification algorithms were proposed with a wide range of estimation performances and complexities [6, 7, 8], but it remains a tough research issue to deploy an adaptive solution for realistic, i.e., noisy and time-varying acoustic environments.

1.2 Signal Processing for Spatial Sound Control

In the reproduction unit of the system in Fig. 1, the signal processing for spatial sound control aims at the presentation of acoustic scenes with at least plausible spatial cues to the users, e.g., the rendering of one or more virtual sound sources at different locations in the acoustic environment. The task of rendering a desired signal at a location of interest exhibits duality with the multichannel acquisition and equalization of point sources using microphone arrays. In both cases, for example, the concept of matched filter arrays has been utilized to achieve the desired focus and selectivity [9, 10]. While the ideal acquisition of a point source, in general, relies on a SIMO (single-input/multiple-output) acoustical model and a subsequent MISO (multiple-input/single-output) inverse processor, the equivalent architecture for the rendering case at hand consists of a SIMO preprocessor and a subsequent MISO acoustic inverse. From the viewpoint of signals and systems, the role of the acoustic impulse responses and the inverse filters are merely exchanged. Essentially, this example shows that the identification of acoustic impulse responses from the loudspeakers to the virtual source location is the key to signal processing for spatial sound control.

Naturally, the rendering of a desired signal at some point of interest will generate the respective soundfield only in the vicinity of the virtual source location, thus not providing the perception of a focused virtual sound source everywhere in the acoustic environment. The point-oriented rendering is underdetermined from the perspective of wavefield synthesis (WFS), which imposes soundfield conditions on a closed contour around the listening area [11]. In order to resolve the underdetermined soundfield conditions, further desired responses can be imposed, e.g., at the locations of the recording microphones in the acoustic environment. The responses in these reference points have to correspond to the soundfield radiated from the desired virtual source. The generation of the desired responses hence requires knowledge of the acoustic system between the virtual source position and the reference points and therefore, again, reliable system identification is indispensable. From the viewpoint of the users in the environment, an excellent set of reference points would be given by the two entrances of the human ear canals.

© EURASIP, 2010 ISSN 2076-1465

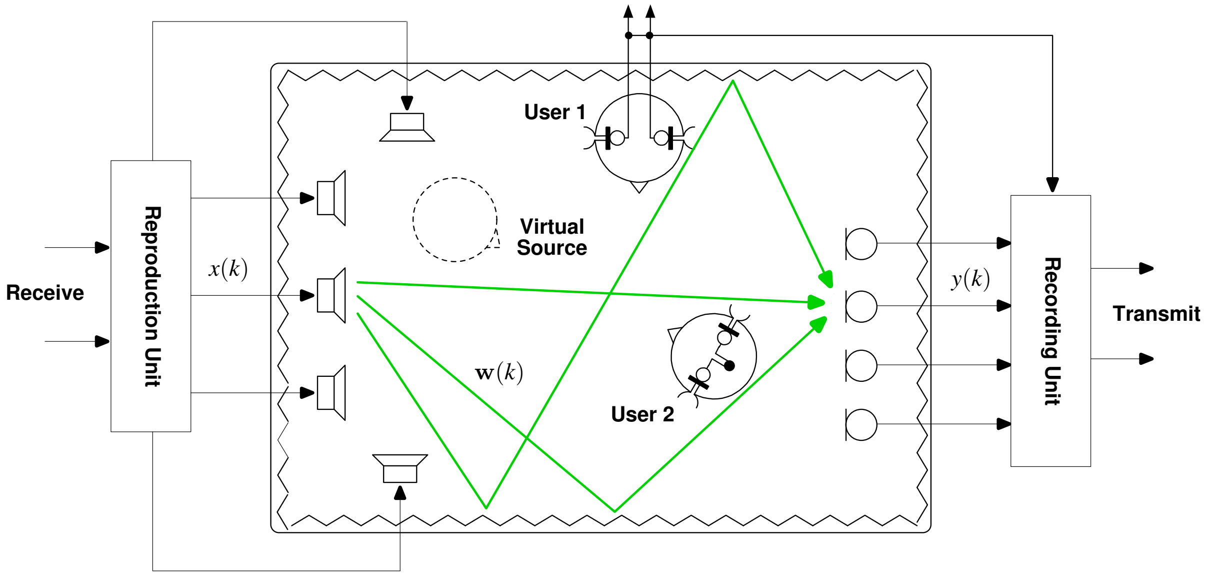

Figure 1: Acoustic environment with loudspeaker input $x(k)$ and microphone output $y(k)$ at discrete-time k. The depicted subset of loudspeaker-to-microphone transmission paths is together represented by the time-varying acoustic impulse response vector $\mathbf{w}(k)$ at time k.

1.3 Acoustic Echo/Interference Cancellation

In case of simultaneous reproduction and recording, we are facing the problem of an undesired feed of the microphones by the loudspeakers of the acoustic environment. Loudspeaker-driven interference at the microphones may severely mislead speech recognition on the recording side of the system, e.g., in speech dialog systems with simultaneous input and output.

Another use case included here is the hands-free voice communication of users ’1’ and ’2’ with a remote acoustic environment, i.e., the destination of the transmit signal in Fig. 1 coincides with the origin of the received signal. A feedback loop might occur only in the worst case, but at least the loudspeaker signals received from the remote-side talkers are superimposed onto the desired speech of user ’1’ and ’2’ at the recording microphones. The feedback to the remote side is then perceived as non-tolerable echo (assuming transmission delay). The undesired loudspeaker-enclosure-microphone system is therefore termed *acoustic echo path* [12].

The acoustic interference in general and the echo signal in particular can be canceled from the recorded microphone signals, while preserving the desired signals, if the time-varying loudspeaker-enclosure-microphone system can be identified accurately from the available signals and utilized for regeneration and subtractive cancellation of the interference [13, 14].

2. SYSTEM IDENTIFICATION TECHNIQUES

As mentioned in the abstract, we will now pick up the particular echo/interference cancellation case with a single active loudspeaker and microphone in order to work out the Bayesian inference model for time-varying acoustic system identification in noise. For structural comparison, we also include the more common deterministic system model and the related estimators. This presentation serves as an example to demonstrate the Bayesian perspective for acoustic system identification, however, not claiming a solution for the variety of system identification problems outlined in the paper.

Consider the single-channel linear time-varying system between the known loudspeaker signal $x(k)$ and the observed microphone signal $y(k)$ as shown by Fig. 1. This configuration provides us with a supervised adaptive system identification problem. Let

$$\mathbf{x}(k) = (x(k), x(k-1), \ldots, x(k-N+1))^T \tag{1}$$

denote an assembly of the most recent input samples and

$$\mathbf{w}(k) = (w_0(k), w_1(k), \ldots w_{N-1}(k))^T \tag{2}$$

the finite set of corresponding impulse response coefficients at sampling time k. Our acoustic system model then reads

$$\begin{aligned} y(k) &= s(k) + \textstyle\sum_{n=0}^{N-1} w_n(k) x(k-n) && (3)\\ &= s(k) + \mathbf{x}^T(k)\mathbf{w}(k)\,, && (4) \end{aligned}$$

where $s(k)$ denotes observation noise, e.g., speech utterances by the users or any kind of ambient noise which originates inside or propagates from the outside into the acoustic environment.

2.1 Deterministic System Model

Given the time-series $\mathbf{y}(k) = (y(k), \ldots, y(k'), \ldots, y(0))$, i.e., observations from the time origin up to and including time k, and furthermore a deterministic and constant model of the unknown system, i.e., $\mathbf{w}(k) = \mathbf{w}$, where $\mathbf{w}$ is now subject to estimation. We can in this case easily formulate the likelihood of the observations as a function of the unknown system, i.e., $p(\mathbf{y}(k) \mid \mathbf{w})$, and target the maximum-likelihood estimate $\widehat{\mathbf{w}}$ of the system as shown by

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}(k) \mid \mathbf{w})\,. \tag{5}$$

By assuming zero-mean, independent, and uncorrelated Gaussian observation noise $s(k)$ with fixed and arbitrary variance σ_s^2, we can write the likelihood over all observations as

$$\begin{aligned} p(\mathbf{y}(k) \mid \mathbf{w}) &= \prod_{k'=0}^{k} \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(y(k') - \mathbf{x}^T(k')\mathbf{w})^2}{2\sigma_s^2}\right)\\ &= \frac{1}{\sqrt{2\pi\sigma_s^2}^k} \exp\left(-\frac{\sum\limits_{k'=0}^{k}(y(k') - \mathbf{x}^T(k')\mathbf{w})^2}{2\sigma_s^2}\right) \end{aligned}$$

and proceed with analytic maximization. For the ease of the formal derivation, this step is typically performed in the logarithmic domain, i.e., the derivative of the log-likelihood $\ln p(\mathbf{y}(k) \mid \mathbf{w})$ with respect to $\mathbf{w}$ is equated to zero. It can be easily verified that this is perfectly in agreement with least-squares estimation of $\mathbf{w}$, which seeks a minimum of the sum of squares of the error signal $e(k) = y(k) - \mathbf{x}^T(k)\mathbf{w}$ by using the best possible selection $\widehat{\mathbf{w}}$ in place of $\mathbf{w}$, i.e.,

$$\widehat{\mathbf{w}} \;=\; \arg\min_{\mathbf{w}} \sum_{k'=0}^{k} \Big(y(k') - \mathbf{x}^T(k')\mathbf{w} \Big)^2 \,, \tag{6}$$

and after basic vector and matrix rearrangements

$$\widehat{\mathbf{w}} \;=\; \left(\sum_{k'=0}^{k} \mathbf{x}(k')\mathbf{x}^T(k') \right)^{-1} \sum_{k'=0}^{k} \mathbf{x}(k') y(k') \,. \tag{7}$$

This resulting estimator is unfortunately rather unpractical, since it causes very high and even growing computational load if the evaluation of the previous equation should happen online for each and every time instant k. This issue can be partly resolved by the recursive least-squares (RLS) algorithm [4], but RLS is still demanding in terms of the resource consumption.

In practice, the gradient descent approach is therefore preferred in many applications. It uses the gradient of the instantaneous squared error, i.e.,

$$\nabla(k) \;=\; \frac{\partial e^2(k)}{\partial \mathbf{w}} \;=\; -2e(k)\mathbf{x}(k) \,, \tag{8}$$

and it achieves the desired smoothing (averaging) in the estimation by introducing a relatively small non-negative step-size factor μ into the recursive prediction and correction mechanism of sequential gradient descent, obtaining the simple and celebrated least mean-square (LMS) adaptive algorithm [4]:

$$\begin{aligned} \widehat{\mathbf{w}}(k+1) \;&=\; \widehat{\mathbf{w}}(k) - \frac{1}{2}\mu\nabla(k) & (9) \\ &=\; \widehat{\mathbf{w}}(k) + \mu e(k)\mathbf{x}(k) \,. & (10) \end{aligned}$$

The described maximum-likelihood (ML) and least-squares (LS) estimators are both rooted in the deterministic model of the acoustic system $\mathbf{w}(k)$. Equations (7) and (10) reveal that this strategy unfortunately does not provide native structural support for the inclusion of comprehensive statistical *a priori* information regarding the unknown system $\mathbf{w}(k)$ or the observation noise $s(k)$. In speech and audio signal processing, however, it has been observed that we ultimately require algorithmic support to handle the very common acoustic scenarios which are stochastic in nature. Particularly, this includes variability of the system $\mathbf{w}(k)$ subject to estimation [14] and in many cases the simultaneous and continuous presence of observation noise $s(k)$ with stationary and non-stationary (i.e., ambient noise and speech-like) characteristics [15, 16].

Before we approach the more sophisticated Bayesian inference model for adaptive system identification, it is acknowledged that research has figured out various auxiliary control mechanisms to support the deterministic learning, e.g., the advanced double talk detection to handle sporadic speech-like observation noise [17], or the time-varying adaptive step-size factors $\mu = \mu(k)$ to provide an optimal balance of tracking ability and robustness against observation noise [12]. Both techniques can be applied to slow down or even halt the adaptive algorithms, e.g., LMS or RLS, in the presence of noise and to accelerate the learning rate otherwise. The plain forms of these adaptive algorithms are, however, not selective with respect to the changing quality of the data and will therefore suffer from performance limitations.

2.2 Proposed Bayesian Inference Model

By employing more stochastic modeling in Bayesian estimation, in contrast to ML and LS, we expect richer and more inherent structural support for robust system identification in time-varying and noisy conditions. Seeking the *posterior* distribution of the unknown system, $p(\mathbf{w}(k) \mid \mathbf{y}(k))$, which can be expressed using Bayes rule, i.e.,

$$p(\mathbf{w}(k) \mid \mathbf{y}(k)) \;=\; \frac{p(\mathbf{y}(k) \mid \mathbf{w}(k))\, p(\mathbf{w}(k))}{p(\mathbf{y}(k))} \,,$$

considerable statistical information about the acoustic environment can be and needs to be incorporated in form of the *likelihood* $p(\mathbf{y}(k) \mid \mathbf{w}(k))$, the *prior* $p(\mathbf{w}(k))$, and the evidence distribution $p(\mathbf{y}(k))$. The obtained *posterior* could then be employed in the context of various estimation criteria as, e.g., *maximum a posteriori (MAP)* estimation or *minimum mean-square error (MMSE)* estimation of the unknown system. MAP estimation is sometimes referred to as poor man's Bayesian estimation, while in the case of jointly Gaussian distributed random variables, the MAP estimator lines up with the MMSE estimator (cf. the previous relationship of maximum-likelihood and least-squares estimation in the Gaussian case). In general, the MMSE estimator of the system $\mathbf{w}(k)$, i.e., the Bayesian estimator under quadratic loss, is analytically given by the conditional mean of $\mathbf{w}(k)$, e.g., [18]:

$$\begin{aligned} \widehat{\mathbf{w}}(k) \;&=\; \mathrm{E}\{\mathbf{w}(k) \,|\, \mathbf{y}(k)\} & (11) \\ &=\; \int_{\mathbf{w}} \mathbf{w}(k)\; p(\mathbf{w}(k) \mid \mathbf{y}(k)) \mathrm{d}\mathbf{w} \,. & (12) \end{aligned}$$

A particularly convenient stochastic model for our time-varying systems $\mathbf{w}(k)$ is the first-order recursive Markov chain, i.e.,

$$\mathbf{w}(k+1) = a \cdot \mathbf{w}(k) + \Delta\mathbf{w}(k) \,, \tag{13}$$

in which the two consecutive states at times k and $k+1$ are related to each other by the transition coefficient $0 \leq a \leq 1$ and the independent process noise quantity $\Delta\mathbf{w}(k)$ with covariance $\boldsymbol{\sigma}^2_\Delta = \mathrm{E}\{\Delta\mathbf{w}(k)\Delta\mathbf{w}^T(k)\}$. The Markov model therefore represents systems which gradually change into an unpredictable direction – very much in agreement with the nature of time-varying impulse responses in realistic acoustic environments.

Equations (4) and (13) together form a Gauss-Markov dynamical model (state-space model) of the unknown state $\mathbf{w}(k)$, provided that we stick to the independent and normally distributed observation and process noises, $s(k)$ and $\Delta\mathbf{w}(k)$, respectively. In this case, the MMSE estimate $\widehat{\mathbf{w}}(k)$ of the unknown system $\mathbf{w}(k)$ at time k, given the observations $y(k)$ up to and including time k, i.e., the *posterior* mean $\widehat{\mathbf{w}}(k) = \mathrm{E}\{\mathbf{w}(k) \,|\, \mathbf{y}(k)\}$, is known to be computed by the Kalman filter which consists of the following set of recursive and iteratively coupled matrix equations [18]:

$$\begin{aligned} \widehat{\mathbf{w}}(k+1) \;&=\; a \cdot \widehat{\mathbf{w}}^+(k) & (14) \\ \mathbf{p}(k+1) \;&=\; a^2 \cdot \mathbf{p}^+(k) + \boldsymbol{\sigma}^2_\Delta & (15) \\ \widehat{\mathbf{w}}^+(k) \;&=\; \widehat{\mathbf{w}}(k) + \mathbf{k}(k)\Big(y(k) - \mathbf{x}^T(k)\widehat{\mathbf{w}}(k) \Big) & (16) \\ \mathbf{p}^+(k) \;&=\; \Big(\mathbf{I} - \mathbf{k}(k)\mathbf{x}^T(k) \Big)\mathbf{p}(k) & (17) \\ \mathbf{k}(k) \;&=\; \mathbf{p}(k)\mathbf{x}(k)\Big(\mathbf{x}^T(k)\mathbf{p}(k)\mathbf{x}(k) + \sigma_s^2(k) \Big)^{-1} . & (18) \end{aligned}$$

Equations (14) and (16) of the Kalman filter recursively determine the conditional mean estimate $\widehat{\mathbf{w}}(k)$ in a prediction-correction fashion. In doing so, the formulas utilize the Kalman gain $\mathbf{k}(k)$ from (18) as a weight which essentially depends on the state error covariance $\mathbf{p}(k)$. The latter is again determined recursively through Eqs. (15) and (17) of the Kalman filter.

The Kalman gain $\mathbf{k}(k)$ can be considered as an intelligent adaptive stepsize parameter of the recursive learning procedure for the system $\widehat{\mathbf{w}}(k)$, comparable to the role of μ in the LMS algorithm, cf. (10). Through the Kalman gain, the model-based "system distance" $\mathbf{p}(k)$ between the true and the estimated acoustic system interacts with the prediction-correction procedure for $\widehat{\mathbf{w}}(k)$. In this way, Kalman filtering can be understood as the ever sought unification of linear adaptive filtering and adaptation control using $\mathbf{p}(k)$. After all, the Kalman filter differs from LMS and RLS by its inherent stability [4], i.e., it does not require additional control mechanisms (e.g., the double-talk detection) in order to achieve fast and yet robust adaptation in time-varying and noisy environments.

For long time, the Kalman filter has been avoided in acoustic system identification. This can be attributed to its still high computational load and to the risk for numerical instability in the case of higher-order adaptive filters [4]. Furthermore, a comprehensive signal model for the Kalman filter, particularly the availability of observation and process noise covariances for the acoustic state-space model in (4) and (13), seemed to be out of sight [12].

2.3 Broadband Kalman Filter

In order to tame the exact Kalman filter, in this paper, we replace the matrix quantity $\mathbf{k}(k)\mathbf{x}^T(k)$ in (17) with the inner vector product $\mathbf{x}^T(k)\mathbf{k}(k)/N$. This seemingly strong simplification can be well justified in the case of broadband input $x(k)$, because the smoothed matrix quantity $\mathbf{k}(k)\mathbf{x}^T(k)$ resembles a near-diagonal correlation matrix, provided that we specify a diagonal process noise covariance $\boldsymbol{\sigma}_\Delta^2 = \sigma_\Delta^2\mathbf{I}$. Along with this replacement, the matrix $\mathbf{p}(k)$ can be treated as a scalar $p(k)$ without further assumption or approximation, as seen from (15) and (17). The normalization by factor N in $\mathbf{x}^T(k)\mathbf{k}(k)/N$ achieves appropriate scaling after the matrix replacement. Because of the broadband rationale behind the rearrangement, the resulting algorithm is termed *broadband Kalman filter*:

$$\widehat{\mathbf{w}}(k+1) = a \cdot \widehat{\mathbf{w}}^+(k) \tag{19}$$

$$p(k+1) = a^2 \cdot p^+(k) + \sigma_\Delta^2 \tag{20}$$

$$e(k) = y(k) - \mathbf{x}^T(k)\widehat{\mathbf{w}}(k) \tag{21}$$

$$\widehat{\mathbf{w}}^+(k) = \widehat{\mathbf{w}}(k) + \mathbf{k}(k)e(k) \tag{22}$$

$$p^+(k) = \Big(1 - \mathbf{x}^T(k)\mathbf{k}(k)/N\Big)p(k) \tag{23}$$

$$\mathbf{k}(k) = p(k)\mathbf{x}(k)\Big(p(k)\mathbf{x}^T(k)\mathbf{x}(k) + \sigma_s^2(k)\Big)^{-1}. \tag{24}$$

By the simplifications introduced here, naturally, the presented algorithm looses its decorrelation ability on the input signal $x(k)$ if non-white input is processed. However, the structural support to handle time-varying unknown systems and to cope with the continuous presence of observation noise with possibly time-varying levels is fully preserved in the *broadband Kalman filter*. Moreover, we have at the same time gained considerable numerical robustness by reducing the dimension of the original estimation error covariance $\mathbf{p}(k)$ from matrix to scalar.

Next, we have to resolve the uncertainty regarding the observation noise power $\sigma_s^2(k)$ in the Kalman gain (24), because this quantity is indispensable for the operation of the Kalman filter. Unfortunately, the corresponding signal $s(k)$ is not available explicitly to calculate sample covariances, but the error signal $e(k)$ in (21) will essentially represent the observation noise signal $s(k)$ in case of successful state estimation. Thus, we can approximate $\sigma_s^2(k) \approx \sigma_e^2(k)$ and then obtain the error signal power $\sigma_e^2(k)$, e.g., by online recursive averaging of the explicitly available square error $e^2(k)$.

Eventually, the scalar process noise covariance parameter required in (20) can be specified as $\sigma_\Delta^2 = (1-a^2)\mathscr{E}\{\mathbf{w}^T(k)\mathbf{w}(k)\}/N$, where $\mathscr{E}\{\mathbf{w}^T(k)\mathbf{w}(k)\}$ denotes the average echo path norm. This formula is deduced directly from the Markov model in (13) when square expectation is applied on both sides. Some insight into the remaining choice of the model parameter a is provided in Sec. 3.

Substituting (22) and (24) into (19), while assuming $\sigma_s^2(k) \approx 0$, the derived algorithm immediately unveils structural equivalence with the normalized LMS algorithm [4], thus proving the numerical efficiency and robustness obtained through simplification of the exact Kalman filter. Our *broadband Kalman filter* with $\sigma_s^2(k) \neq 0$ in fact represents an excellent tradeoff between the Bayesian inference in form of the exact Kalman filter and the very popular LMS-type adaptive algorithm for acoustic system identification.

3. EXPERIMENTAL COMPARISON

Let us again consider the echo/interference cancellation case from Sec. 1.3. The somewhat harsh, but reproducible configuration in our evaluation uses time-varying echo paths $\mathbf{w}(k)$ conforming to the Markov model (13) with time-constant $\tau_w = -1/(f_s \ln(a))$. The audio sampling frequency used here is $f_s = 8\,kHz$. $\mathscr{E}\{\mathbf{w}^T(k)\mathbf{w}(k)\}$ is unity and the average echo-to-near-end speech power ratio thus 0 dB. The echo path has 500 coefficients with exponential decay characteristics. Simultaneous with different degrees of echo path variation, i.e., different τ_w, we consider a double-talk condition with continuous presence of both far-end and near-end speech, $x(k)$ and $s(k)$, respectively. Fig. 2 depicts the corresponding echo signal at the recording microphone and the additive near-end speech.

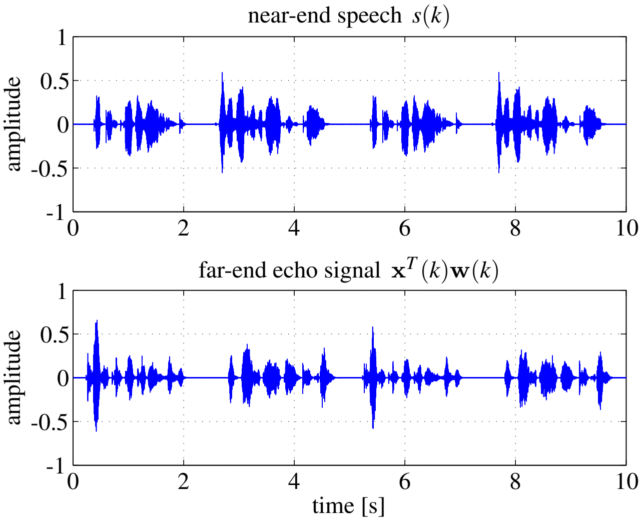

Figure 2: Signals in echo/interference cancellation configuration.

Providing for similar computational complexities of different algorithms in the experiment, we consider on the one hand the popular NLMS algorithm, i.e., Eq. (10) with $\mu = \mu_0/(\mathbf{x}^T(k)\mathbf{x}(k) + \delta)$, $\delta = 0.02$, and on the other hand the *broadband Kalman filter* derived in Sec. 2.3 of this paper. In both cases, the adaptive filter length is chosen as $N = 300$. Systems with adaptive stepsize control based on double-talk detection are not an option for comparison, here, because of the permanent double-talk condition. For the operation of the *broadband Kalman filter*, we choose the model parameter $a = 0.999984$ in all simulations. The resulting process noise covariance σ_Δ^2 in the Kalman filter then ideally suits an unknown system with time-constant $\tau_w \approx 8\,s$.

Performance comparison of the time-varying system identification is achieved in terms of the normalized system distance $D(k) = (\mathbf{w}(k) - \widehat{\mathbf{w}}(k))^T(\mathbf{w}(k) - \widehat{\mathbf{w}}(k))/(\mathbf{w}^T(k)\mathbf{w}(k))$ at each and every time k. While using the same algorithms, Figs. 3 and 4 present the results obtained for time-varying and almost time-invariant echo paths, i.e., for model match and model mismatch with respect to the Kalman filter. For reference, we included fast and slow variants of the NLMS algorithm using different adaptation constants μ_0. It can be seen that the system distance for the fast NLMS is erratic in these harsh noisy conditions imposed by the near-end speaker, allowing only between 0 and -5 dB most of the times for the time-varying system in Fig. 3. The slower NLMS variant behaves less erratic, but it looses track of the time-varying system and then sometimes the system distance even increases. The *broadband Kalman filter*, on the contrary, initially converges much more rapidly than the NLMS and then tracks the time-varying system with an average system distance in the range of -10 dB with natural ups and downs according to the strongly time-varying conditions in this experiment. The results in Fig. 4 for the mismatched, almost time-invariant echo path are generally better. This is due to the naturally better identifiability of slowly varying systems by whatever algorithm. Thereby, we observe stronger advantages for the broadband Kalman filter.

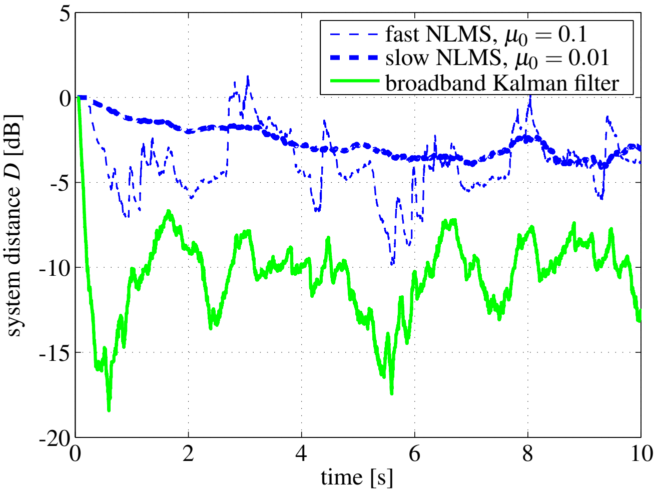

Figure 3: Results for time-varying system: $\tau_w \approx 8\ s$ (model match).

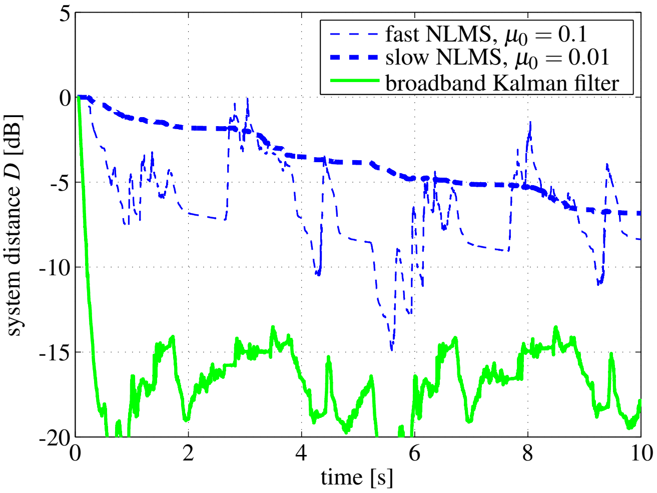

Figure 4: Time-invariant system, i.e., $\tau_w \longrightarrow \infty$ (model mismatch).

4. CONCLUSIONS

Many applications of acoustic signal processing can be handled successfully with the availability of the impulse responses of the acoustic environment, but the large variety of possible loudspeaker and microphone setups and the various reflection characteristics do not suggest the use of predetermined impulse responses in most of the use cases. The crux further lies in the time-varying nature of the acoustic environment (e.g., by user interaction) and the simultaneous presence of observation noise (e.g., ambient noise or competing speech sources). As a result, the uncertainty regarding the current state of a time-varying acoustic system needs to be resolved by fast and robust online system identification.

In order to take the stochastic behavior of acoustic environments into account, this paper suggested a Bayesian inference model for acoustic system identification. In contrast to deterministic modeling and the related ML/LS/RLS estimation, the Bayesian inference provides richer support for the inclusion of *a priori* information, e.g., regarding the variability of the environment. For the supervised system identification case, it was finally demonstrated how the Bayesian inference can be handled to provide us with simple and robust adaptive algorithms for adverse conditions.

REFERENCES

[1] Michael Brandstein and Darran Ward, *Microphone Arrays*, Springer, 2001.

[2] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing*, Springer, 2008.

[3] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, February 1988.

[4] Simon Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, 4th edition, 2002.

[5] K. Abed-Meraim, W. Qui, and Y. Hua, “Blind system identification,” *in Proc. of the IEEE*, vol. 85, no. 8, pp. 1310–1322, August 1997.

[6] Y. Huang and J. Benesty, “Adaptive multi-channel least mean-square and Newton algorithms for blind channel identification,” *Signal Process.*, , no. 82, pp. 1127–1138, 2002.

[7] S. Gannot and M. Moonen, “Subspace methods for multimicrophone speech dereverberation,” *EURASIP Journal Applied Signal Process.*, , no. 11, pp. 1074–1090, 2003.

[8] D. Schmid and G. Enzner, “Robust subsystems for iterative multichannel blind system identification and equalization,” *in Proc. 2009 IEEE Pacific Rim Conf. on Commun., Comput. and Signal Process.*, pp. 889–893, August 2009.

[9] E.-E. Jan, P. Svaizer, and J.L. Flanagan, “Matched-filter processing of microphone array for spatial volume selectivity,” in *IEEE International Symposium on Circuits and Systems*, May 1995, pp. 1460–1463.

[10] Sylvain Yon, Mickael Tanter, and Mathias Fink, “Sound focusing in rooms: The time-reversal approach,” *Journal of the Acoustical Society of America*, vol. 113, no. 3, pp. 1533–1543, March 2003.

[11] A. Berkhout, D. de Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[12] Eberhard Hänsler and Gerhard Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley, 2004.

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

[14] Gerald Enzner and Peter Vary, “Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones,” *Signal Processing, Elsevier*, vol. 86, no. 6, pp. 1140–1156, June 2006.

[15] Toon van Waterschoot, Geert Rombouts, Piet Verhoeve, and Marc Moonen, “Double-talk robust prediction error identification algorithms for acoustic echo control,” *IEEE Trans. on Signal Processing*, vol. 5, pp. 846–858, March 2007.

[16] S. Malik and G. Enzner, “Model-based vs. traditional frequency-domain adaptive filtering in the presence of continuous double-talk and acoustic echo path variability,” in *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, September 2008.

[17] J. Benesty, D.R. Morgan, and J.H. Cho, “A new class of double talk detectors based on cross-correlation,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 168–172, March 2000.

[18] Louis L. Scharf, *Statistical Signal Processing*, Addison-Wesley Publishing Company, 1991.