Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise

Doire, Clement Samuel Joseph; Brookes, Mike; Naylor, Patrick A.; Hicks, Christopher M.; Betts, Dave; Dmour, Mohammad A.; Jensen, Soren Holdt

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2016.2641904

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise

Clement S. J. Doire, *Student Member, IEEE*, Mike Brookes, *Member, IEEE*, Patrick A. Naylor, *Senior Member, IEEE*, Christopher M. Hicks, Dave Betts, Mohammad A. Dmour, *Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—This paper proposes an online single-channel speech enhancement method designed to improve the quality of speech degraded by reverberation and noise. Based on an autoregressive model for the reverberation power and on a hidden Markov model for clean speech production, a Bayesian filtering formulation of the problem is derived and online joint estimation of the acoustic parameters and mean speech, reverberation, and noise powers is obtained in mel-frequency bands. From these estimates, a real-valued spectral gain is derived and spectral enhancement is applied in the short-time Fourier transform (STFT) domain. The method yields state-of-the-art performance and greatly reduces the effects of reverberation and noise while improving speech quality and preserving speech intelligibility in challenging acoustic environments.**

***Index Terms*—Dereverberation, speech, Bayesian, single-channel.**

I. Introduction

SPEECH signals captured using a distant microphone within a confined acoustic space are often corrupted by reverberation. The detrimental impact of reverberation on the quality and intelligibility of the speech and on the performance of speech recognition systems is made worse when it is combined with acoustic noise [1]–[4]. Combating the damaging effects of reverberation has been a key research topic in recent years driven by an increasing demand for effective methods of speech communication in challenging environments [5]. While some progress has been made in both single- and multi-channel processing [4], [6]–[9], the task of providing a blind single-channel dereverberation method robust to noise and suitable for real-time processing remains a challenge.

Most single-channel speech dereverberation techniques can be classified into inverse filtering [10], [11], nonlinear mapping [12], spectral enhancement [6], [13], [14] and probabilistic model-based methods [15]–[17]. Inverse filtering methods typically try to reconstruct the original signal by designing an inverse filter for the Room Impulse Response (RIR). Based on the observation that the Linear Prediction (LP) residual of clean speech has a higher kurtosis (fourth-order moment) than that of reverberant speech, the inverse filter of the impulse response is estimated in [10] by maximizing the kurtosis of the LP residual of the inverse-filtered speech. In [11], a similar principle is applied, in which the inverse filter is chosen to maximize the normalized skewness (third-order moment) of the LP residual. These techniques, however, compensate only for the coloration effect caused by the early reflections and must be used in conjunction with other late reverberation suppression methods in order to achieve good dereverberation performance [10], [11]. If the RIR is known, or can be estimated, inverse filtering can also be applied using methods in the time or frequency domain [18] or using homomorphic approaches [19], [20].

Nonlinear mapping methods do not assume any explicit model for the reverberation, and instead use parallel training data in order to learn a nonlinear mapping function from the reverberant speech spectrogram to its clean speech equivalent. This can be done using a fully connected Deep Neural Network (DNN) as in [12] where the mean squared error between the output of the DNN and the clean speech log-power spectrum is minimized. Even though results can be improved by also considering first and second-order time derivatives of the input features, speech enhanced by this method can lead to a decrease in overall speech quality [21].

In spectral enhancement methods, a time-frequency gain is applied to the noisy reverberant spectral coefficients in order to estimate those of the clean speech. This gain is based on the estimated power spectral densities (PSDs) of the noise and late reverberation [6], [13]. The estimation of the late reverberant PSD is often based on a simple statistical model of the room impulse response such as [6], [22]. Spectral enhancement methods are able to reduce both the background noise and reverberation while being computationally efficient, but usually suffer from artifacts introduced by the nonlinear filtering operation, though efforts have been made to alleviate this problem, e.g. by using temporal cepstrum smoothing [14].

In the probabilistic model-based approaches to blind dereverberation, the parameters of the acoustic channel and clean

Manuscript received July 1, 2016; revised November 3, 2016 and December 12, 2016; accepted December 15, 2016. Date of publication December 22, 2016; date of current version January 26, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard Christian Hendriks.

C. S. J. Doire, M. Brookes, and P. A. Naylor are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: clement.doire@gmail.com; mike.brookes@imperial.ac.uk; p.naylor@imperial.ac.uk).

C. M. Hicks, D. Betts, and M. A. Dmour are with the CEDAR Audio, Ltd., Cambridge CB21 5BS, U.K. (e-mail: christopher.hicks@cedaraudio.com; dave.betts@cedaraudio.com; mohammad.dmour@cedaraudio.com).

S. H. Jensen is with the Department of Electronic Systems, Aalborg University, Aalborg 9100, Denmark (e-mail: shj@es.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2641904

speech models are estimated from the observed data and used to reconstruct the original source signal. The reverberation model is typically an FIR or IIR filter in the time domain [15], the complex short-time Fourier transform (STFT) domain [23], [24] or the STFT power domain [16]. In [15], the acoustic channel is modeled as a time-varying linear combination of entries from a codebook of all-pole filters, and the speech signal is modeled using a block-based time-varying autoregressive (AR) model. Bayesian inference is used to estimate the joint probability density function (pdf) of the channel and source parameters. The method has been applied successfully on simulated data within a limited frequency range, but difficulties arise when the data does not follow the assumed channel and source models. Bayesian variational inference is used in [16] where an extension of the Multi-Channel Linear Prediction (MCLP) model [25] to power spectrograms in the single-channel case is used. The order of this non-negative auto-regressive reverberation model is determined in a data-driven manner using a Dirichlet process [26]. However, the method assumes a noise-free environment, which is unrealistic in practice. In [17], a Non-negative Convolutive Transfer Function (N-CTF) model [8] is used for the RIR and the speech spectrogram is modeled using Non-negative Matrix Factorization (NMF) so as to capture the spectral structure of the speech signal. The two models are then combined to form an optimization problem in which the clean speech spectrogram and RIR parameters are simultaneously estimated through iterative update rules. In [24], the reverberation model is an FIR filter in the complex STFT domain. Processing each subband independently, a recursive expectation-maximization (EM) procedure is used in which the E step estimates the clean-speech coefficients with a Kalman filter and the M step updates a parameter vector comprising the reverberation filter coefficients and the variances of the speech and noise.

In this paper, we present an online method for enhancing reverberant and noisy speech recordings using a combination of spectral enhancement and probabilistic estimation. Enhancement is performed by applying a time-frequency gain to the degraded speech complex STFT coefficients as in spectral enhancement. The estimation of the quantities needed to compute this gain is formulated as a Bayesian filtering problem in which they are jointly estimated along with the parameters of the acoustic channel. The latter is modeled using a non-negative first-order autoregressive moving-average (ARMA) process parametrized by the reverberation time (T_{60}) and the Direct-to-Reverberant energy Ratio (DRR). The clean speech is modeled by a Hidden Markov Model (HMM) in which each state captures the spectral characteristics of a possible prior distribution of the multivariate speech log-power. At each time frame, the possible clean speech prior distributions are tested through a swarm of nonlinear Kalman filter-like updates. The distribution leading to the highest likelihood for the observed power is kept, leading to a-posteriori estimates of the speech, reverberation and noise mean powers. The performance of the proposed method is evaluated on simulated data through six different objective measures and on live recordings through the Word Error Rate (WER) of a speech recognizer. A listening test was conducted to assess the subjective reverberation reduction and overall quality improvement. The idea of using an HMM whose states represent broad speech sound classes with distinct acoustic spectra has been applied previously to speech enhancement [27]–[30]. In these papers, a state-dependent spectral shape was multiplied by a time-varying speech gain to obtain prior distributions for the speech spectral coefficients; these priors were then used to determine an MMSE or MAP estimate of the clean speech spectrum in an appropriate domain. In the current work, this approach is extended to include an explicit model of reverberation and to track the time-variation of both the reverberation model parameters and the speech gain.

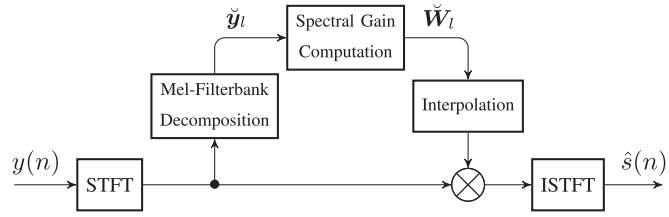


Fig. 1. Enhancement system overview.

The paper is organized as follows. The non-negative ARMA reverberation model and HMM clean speech model are described in Section II and an overview of the overall enhancement system is given in Section III. In Section IV, the Bayesian filtering formulation of the problem is detailed as well as the computation of the posterior densities and the online estimation of the reverberation parameters. Results are presented in Section V and conclusions drawn in Section VI.

II. Signal Model and Notation

In the system block diagram shown in Fig. 1, the enhancement of the noisy and reverberant speech is performed in the STFT domain, while the estimation of the system parameters and signal powers is performed in Mel-spaced subbands. A filterbank comprising triangular filters [31], [32] is used to transform the power spectrum of each frame from $\tilde{K}$ STFT bins to a reduced number, K, of Mel-spaced subbands. The use of these broad subbands has two benefits: it reduces the dimension of the state vector, $\boldsymbol{x}_l$, in (14) below and it reduces the number of states required in the speech model described in Section II-B. This is because the filterbank removes narrowband features such as pitch harmonics whose variability would otherwise need to be included in the model.

A. Reverberation Model

Let $y(n)$ denote the observed reverberant noisy speech signal at discrete-time n. The additive background noise signal is denoted by $\nu(n)$ and the reverberant speech signal is obtained by convolving the clean speech source $s(n)$ and the J-tap RIR between the source and microphone, $\rho(n)$, as

$$y(n) = \sum_{r=0}^{J-1} \rho(r) s(n-r) + \nu(n). \tag{1}$$

The complex STFT coefficients of the observed signal are then computed according to

$$Y^{\circ}(l,\tilde{k}) = \sum_{n=0}^{\tilde{K}-1} y(n+lT)w(n)e^{-j\frac{2\pi}{\tilde{K}}n\tilde{k}} \tag{2}$$

where l is the time-frame index, $\tilde{k}$ is the STFT frequency bin, $w(n)$ a time-domain window, and T the frame increment. A power-domain filterbank is applied to compute the power in K Mel-spaced subbands as

$$\breve{Y}(l,k) = \sum_{\tilde{k}=0}^{\tilde{K}-1} b_{k,\tilde{k}} |Y^{\circ}(l,\tilde{k})|^2. \tag{3}$$

where the $b_{k,\tilde{k}}$ implement the triangular filters from [31], [32]. Analogous to (3), $\breve{N}(l,k)$ denotes the subband noise power. For the speech signal, however, we divide by the band-independent active speech level [33], $\breve{G}(l)$, to obtain the level-normalized subband speech signal

$$\breve{S}(l,k) = \frac{1}{\breve{G}(l)} \sum_{\tilde{k}=0}^{\tilde{K}-1} b_{k,\tilde{k}} |S^{\circ}(l,\tilde{k})|^2. \tag{4}$$

The decomposition of the speech power into the product of a time-varying active level, $\breve{G}(l)$, and a level-normalized spectral shape, $\breve{S}(l,k)$, is similar to that in [28], [30] and allows the prior distribution of $\breve{S}(l,k)$ to be trained offline using level-normalized training data.

Based on an approximation of (1) in the STFT domain in which cross-band filters are neglected, the N-CTF model was proposed in [8] to approximate the power spectrogram of a reverberant signal. In this paper, we assume this model to apply in each Mel-frequency band, k, and also assume that the reverberant speech and noise are additive in the power domain, resulting in

$$\breve{Y}(l,k) = \sum_{\tau=0}^{L_h-1} \breve{H}(\tau,k)\breve{G}(l-\tau)\,\breve{S}(l-\tau,k) + \breve{N}(l,k) \tag{5}$$

where L_h is the RIR length in the STFT domain. The errors introduced by assuming that the signals add in the power domain are discussed further in Appendix B.

Polack proposed a time-domain statistical model [22] of the RIR as scaled exponentially-decaying white Gaussian noise parametrized by the broadband reverberation time T_{60}. Noting that the latter is normally frequency-dependent [34], this model was extended in [6] to each subband and split into two statistical sub-models: one containing the direct path, the other comprising all later reflections. In this paper, we assume the exponentially-decaying model is valid in each Mel-frequency band, model the direct path deterministically and only consider the energy envelope of the impulse response so that

$$\breve{H}(l,k) = \delta(l) + d_k\,\alpha_k^{l-1} u(l-1) \tag{6}$$

where $\delta(l)$ is the Kronecker delta function and $u(l)$ is the unit step function. The decay constant, α_k, in Mel-frequency band k, is related to $T_{60,k}$ through

$$\alpha_k^{\frac{T_{60,k}}{T}} = 10^{-6} \tag{7}$$

where T is the STFT frame hop. The drop in energy after the direct path, d_k, is related to the frequency-dependent DRR by the equation

$$d_k = \frac{1-\alpha_k}{\mathrm{DRR}_k}. \tag{8}$$

Substituting the drop and decay reverberation model of (6) into the observed power of (5), we obtain

$$\breve{Y}(l,k) = \breve{G}(l)\,\breve{S}(l,k) + \breve{R}(l,k) + \breve{N}(l,k). \tag{9}$$

where the reverberation power in time-frame l and frequency band k is

$$\breve{R}(l,k) = \sum_{\tau=1}^{L_h} d_k \breve{G}(l-\tau)\breve{S}(l-\tau,k)\alpha_k^{\tau-1}. \tag{10}$$

The model of (10) can be written recursively as

$$\breve{R}(l,k) = d_k\,\breve{G}(l-1)\breve{S}(l-1,k) + \alpha_k\,\breve{R}(l-1,k). \tag{11}$$

Equations (9) and (11) correspond to a first-order ARMA model for the acoustic channel in the spectral power domain having the system function $\frac{z-\alpha+d}{z-\alpha}$. This parsimonious model contrasts with the higher order moving average or autoregressive models used by [23] and [24] respectively in the complex STFT domain. By writing the frequency-dependent quantities in (9) and (11) as column vectors of length K, we can write the system's dynamic equations as

$$\breve{\boldsymbol{R}}_l = \boldsymbol{\alpha}_{l-1} \odot \breve{\boldsymbol{R}}_{l-1} + \boldsymbol{d}_{l-1} \odot \breve{G}_{l-1}\breve{\boldsymbol{S}}_{l-1} \tag{12}$$

$$\breve{\boldsymbol{Y}}_l = \breve{G}_l\breve{\boldsymbol{S}}_l + \breve{\boldsymbol{R}}_l + \breve{\boldsymbol{N}}_l \tag{13}$$

where $\odot$ is the Hadamard product. In the following, uppercase letters represent random variables, the corresponding lower case letters their realizations, and estimates are denoted by $\hat{}$. Means and covariances are denoted by μ, Σ with the random variable as a suffix. Unadorned signal variables are in the log-power domain and the corresponding power domain quantities indicated by a $\breve{}$; thus $\boldsymbol{y}_l = \log(\breve{\boldsymbol{y}}_l)$. A sequence of consecutive frames is represented using a colon; thus $\boldsymbol{y}_{1:l}$ denotes $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_l\}$. We assume below that the log-power spectra, $\boldsymbol{S}_l$, $\boldsymbol{R}_l$ and $\boldsymbol{N}_l$, follow multivariate Gaussian distributions [35].

B. Clean Speech Model

The log-power, $\boldsymbol{S}_l$, of the level-normalized clean speech is modeled by an HMM with N states in which the state at time frame l is denoted by c_l. Associated with each state is a prior distribution for the multivariate clean speech log-power, so that $p(\boldsymbol{S}_l|c_l) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}}, \boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}})$ where the $\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}$ are trained offline using the training procedure discussed in Section V-A1.

We denote by $\mathbf{c}_l$ the path, $\{c_1, c_2, \ldots, c_l\}$ ending in c_l. For each possible state, c_l, at time frame l, we consider the N possible paths $\{\mathbf{c}_{l-1}, c_l\}$ and select the one with highest likelihood

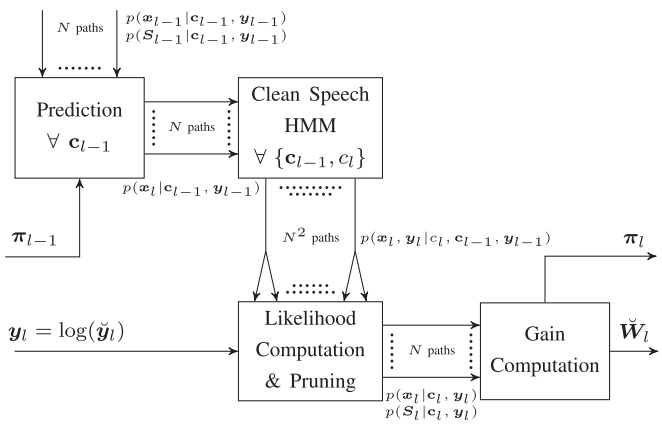


Fig. 2. Bayesian gain computation system described in Section IV.

as $\mathbf{c}_l$ (see (25) below). Thus, for each time frame, we end up with N hypothesized paths, $\mathbf{c}_l$, one for each of the N states.

III. System Overview

To perform enhancement, the reverberant noisy speech signal, $y(n)$, is processed by applying a real-valued magnitude gain to its complex STFT coefficients in order to obtain the estimated clean speech signal $\hat{s}(n)$. This gain is first computed in each Mel-frequency band at each time-frame and then interpolated to cover the full STFT frequency range, as illustrated in Fig. 1.

A. Clean Speech HMM

As we want to track the system parameters over time, the computation of the spectral gain in the Mel-frequency bands, shown as the upper block in Fig. 1, uses a Bayesian filtering formulation that is illustrated in Fig. 2. This includes the clean speech HMM which encapsulates prior speech knowledge in the form of state transition probabilities and state-dependent log-power spectral distributions.

We define

$$\boldsymbol{x}_l = (G_l, \boldsymbol{R}_l, \boldsymbol{N}_l)^T\ , \tag{14}$$

of size $2K+1$, to be the state representation of our system at frame l. Note that $\boldsymbol{x}_l$ includes the reverberation and noise parameters for all subbands in a single state vector in contrast to algorithms such as [23], [24] in which each subband is processed independently. The inclusion of all subbands in a single state vector enables our algorithm to take account of inter-band correlations of the reverberation and noise parameters.

For each of the N best paths, $\mathbf{c}_{l-1}$, the "Prediction" block in Fig. 2 estimates the prior distribution of $\boldsymbol{x}_l$ from the path-dependent posterior distributions of $\boldsymbol{x}_{l-1}$ and $\boldsymbol{S}_{l-1}$. To do so, it uses the current estimate of the reverberation parameters contained in the vector $\boldsymbol{\pi}_{l-1}$ (defined fully in (53) below). For each of these paths, N new possibilities arise, corresponding to the possible prior distributions for the clean speech log-power associated with each HMM state c_l. This gives N^2 possible likelihood functions for the observed log-power $\boldsymbol{y}_l$, corresponding to the N^2 possible choices $\{\mathbf{c}_{l-1}, c_l\}$. The "Likelihood Computation & Pruning" block then computes the likelihood of each of the N^2 paths. Only the path arriving at each c_l with the highest likelihood is kept, and new path-dependent posterior distributions for $\boldsymbol{x}_l$ and $\boldsymbol{S}_l$ are computed, as described in Section IV-B2.

B. Gain Computation

For each time-frame l, we obtain the Gaussian posterior densities of the state vector $\boldsymbol{x}_l$ and clean speech log-power $\boldsymbol{S}_l$ conditional on the HMM path, c_l as described in Section IV-B. From these, an updated estimate for the reverberation parameters $\boldsymbol{\pi}_l$ is computed as described in Section IV-C. The path probabilities, $p(c_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{1:l})$, are normally extremely sparse in practice; in the final block of Fig. 2, we therefore compute the speech enhancement gain, $\breve{\boldsymbol{W}}_l$, from the posterior pdfs of the clean speech, reverberation and noise log-powers associated with the most probable path. From the mean and covariance of the distribution in the log-power domain, we obtain the mean of the corresponding distribution in the power domain using the formulae relating the moments of a normal distribution in the log-power domain to those of a log-normal distribution in the power domain [36]:

$$\boldsymbol{\mu}_{\breve{\boldsymbol{x}}_l} = \exp\left(\boldsymbol{\mu}_{\boldsymbol{x}_l} + \frac{1}{2}\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{x}_l})\right) \tag{15}$$

where $\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{x}_l})$ is the vector composed of the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{x}_l}$. Similarly, we can obtain $\boldsymbol{\mu}_{\breve{\boldsymbol{S}}_l}$ from the mean and covariance matrix of its log-domain distribution. We can then directly extract the estimated means of $\breve{\boldsymbol{R}}_l$, $\breve{\boldsymbol{N}}_l$, $\breve{G}_l$ and $\breve{\boldsymbol{S}}_l$.

According to (13) we have $\breve{\boldsymbol{Y}}_l = \breve{G}_l\breve{\boldsymbol{S}}_l + \breve{\boldsymbol{R}}_l + \breve{\boldsymbol{N}}_l$, and we wish to compute an estimate of the clean speech power $\breve{G}_l\breve{\boldsymbol{S}}_l$ as

$$\widehat{\breve{G}_l\breve{\boldsymbol{S}}_l} = \breve{\boldsymbol{W}}_l^2 \odot \breve{\boldsymbol{Y}}_l \tag{16}$$

where $\breve{\boldsymbol{W}}_l$ is a magnitude gain. This is a form of spectral subtraction [37], [38], and a general form for the gain $\breve{\boldsymbol{W}}_l$ is

$$\breve{\boldsymbol{W}}_l = \left(\frac{\boldsymbol{\mu}_{\breve{G}_l}\boldsymbol{\mu}_{\breve{\boldsymbol{S}}_l}}{\boldsymbol{\mu}_{\breve{G}_l}\boldsymbol{\mu}_{\breve{\boldsymbol{S}}_l} + \eta(\boldsymbol{\mu}_{\breve{\boldsymbol{R}}_l} + \boldsymbol{\mu}_{\breve{\boldsymbol{N}}_l})}\right)^{\beta} \tag{17}$$

where the division and power operations act elementwise on the vectors. η is the oversubtraction factor, and controls how aggressively the processing is applied. Depending on the value of the exponent β, several forms of spectral enhancement can be obtained. The value of β determines the sharpness of the transition from $\breve{\boldsymbol{W}}_l(k) = 1$ to $\breve{\boldsymbol{W}}_l(k) = 0$ [39], with $\beta = 1$ (corresponding to Wiener-Filtering) achieving more aggressive processing than $\beta = \frac{1}{2}$.

Since the estimation of the posterior density of $\breve{\boldsymbol{S}}_l$ is based on a discrete choice of priors at each time-frame, the resulting estimated $\boldsymbol{\mu}_{\breve{\boldsymbol{S}}_l}$ is highly time varying. Accordingly, we perform smoothing of the gain in the time domain according to

$$\breve{\boldsymbol{W}}_l = \lambda_s\breve{\boldsymbol{W}}_{l-1} + (1-\lambda_s)\left(\frac{\boldsymbol{\mu}_{\breve{G}_l}\boldsymbol{\mu}_{\breve{\boldsymbol{S}}_l}}{\boldsymbol{\mu}_{\breve{G}_l}\boldsymbol{\mu}_{\breve{\boldsymbol{S}}_l} + \eta(\boldsymbol{\mu}_{\breve{\boldsymbol{R}}_l} + \boldsymbol{\mu}_{\breve{\boldsymbol{N}}_l})}\right)^{\beta} \tag{18}$$

where λ_s is the smoothing constant. Finally, as indicated in Fig. 1, we use linear interpolation to map the gain, $\breve{\boldsymbol{W}}_l$, from K Mel-spaced bands onto the full STFT resolution. The effect

of this interpolation is to smooth the gain function in frequency, which helps to reduce artifacts such as musical noise.

IV. Bayesian Estimation

In this section, we are concerned with the computation of the posterior densities of the state vector $\boldsymbol{x}_l$ and clean speech log-power $\boldsymbol{S}_l$ in order to be able to perform the gain computation described in Section III-B. The general structure of the proposed algorithm is illustrated in Fig. 2 and detailed in Section IV-A. Section IV-B describes the computation of the means and covariance matrices of the Gaussian pdfs involved, while Section IV-C details how to update the reverberation parameters estimate.

We denote by $\boldsymbol{\mu}_{\boldsymbol{x}_l}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}_l}$ the mean and covariance matrix of the probability density function of $\boldsymbol{x}_l$. Given c_l, the HMM state at time l, we have available from the training data and as detailed in Section II-B the corresponding mean $\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}$ of the prior distribution $p(\boldsymbol{S}_l|c_l)$.

We can describe our system dynamics with the following nonlinear prediction and observation equations:

$$\boldsymbol{x}_l = f(\boldsymbol{x}_{l-1}, \boldsymbol{S}_{l-1}) + \boldsymbol{\epsilon}_l \tag{19}$$

$$\boldsymbol{y}_l = h(\boldsymbol{x}_l, \boldsymbol{S}_l) + \boldsymbol{\nu}_l \tag{20}$$

in which $\boldsymbol{\epsilon}_l \sim \mathcal{N}(0, \boldsymbol{Q}_l)$ and $\boldsymbol{\nu}_l \sim \mathcal{N}(0, \boldsymbol{M}_l)$. The function $f : \mathbb{R}^{3K+1} \rightarrow \mathbb{R}^{2K+1}$ in (19) implements (12) as

$$\begin{cases} G_l = G_{l-1} \\ \boldsymbol{R}_l = \log\left(\boldsymbol{\alpha}_{l-1} \odot \exp\left(\boldsymbol{R}_{l-1}\right) + \boldsymbol{d}_{l-1} \odot \exp\left(G_{l-1} + \boldsymbol{S}_{l-1}\right)\right) \\ \boldsymbol{N}_l = \boldsymbol{N}_{l-1} \end{cases} \tag{21}$$

where $\boldsymbol{\alpha}_{l-1}$ and $\boldsymbol{d}_{l-1}$ are assumed to be system parameters, fixed for time-frame l. From (21), we see that the speech gain, G_l, and the noise log-power, $\boldsymbol{N}_l$, follow a Gaussian random walk. The function $h : \mathbb{R}^{3K+1} \rightarrow \mathbb{R}^K$ in (20) implements (13) as $h(\boldsymbol{x}_l, \boldsymbol{S}_l) = \log\left(\exp(G_l + \boldsymbol{S}_l) + \exp(\boldsymbol{R}_l) + \exp(\boldsymbol{N}_l)\right)$. The nonlinear functions f and h are both differentiable as required for implementing the extended Kalman filter update described in Section IV-B below. The covariance of $\boldsymbol{\epsilon}_l$ is $\boldsymbol{Q}_l$ and represents the variance of the errors in the prediction model, (21). Similarly, $\boldsymbol{M}_l$, the observation noise covariance, represents both the errors inherent to the statistical properties of the input data and those introduced by assuming that uncorrelated signals add in the power domain; expressions for the two components of $\boldsymbol{M}_l$ are derived in Appendices A and B respectively.

From our system equations, we can derive several conditional independencies. Given $\boldsymbol{x}_l$ and c_l, $p(\boldsymbol{y}_l|\boldsymbol{x}_l, c_l, \boldsymbol{y}_{1:l-1}) = p(\boldsymbol{y}_l|\boldsymbol{x}_l, c_l)$. We also have $P(c_l|c_{l-1}, \boldsymbol{x}_{l-1}, \boldsymbol{y}_{1:l-1}) = P(c_l|c_{l-1})$ using pre-trained transition probabilities.

A. State Sequence Estimation

We want to maximize the joint likelihood of the path through the HMM and the sequence of observations, marginalizing over the system state $\boldsymbol{x}_l$. Assume we know $p(\boldsymbol{y}_{1:l-1}, \mathbf{c}_{l-1})$, the probability of a path up until time $l-1$, as well as the posterior density functions $p(\boldsymbol{x}_{l-1}|\boldsymbol{y}_{l-1}, \mathbf{c}_{l-1})$ and $p(\boldsymbol{S}_{l-1}|\boldsymbol{y}_{l-1}, \mathbf{c}_{l-1})$. We can compute:

$$p(\boldsymbol{y}_{1:l}, \mathbf{c}_l) = p(\boldsymbol{y}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_{l-1}) P(c_l|c_{l-1}) p(\boldsymbol{y}_{1:l-1}, \mathbf{c}_{l-1}) \tag{22}$$

where

$$p(\boldsymbol{y}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_{l-1}) = \int_{\boldsymbol{x}_l} p(\boldsymbol{y}_l|c_l, \boldsymbol{x}_l) p(\boldsymbol{x}_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{l-1}) d\boldsymbol{x}_l \tag{23}$$

in which

$$p(\boldsymbol{x}_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{l-1}) = \int_{\boldsymbol{x}_{l-1}} p(\boldsymbol{x}_{l-1}, \boldsymbol{x}_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{l-1}) d\boldsymbol{x}_{l-1}. \tag{24}$$

For each of the N possible $\mathbf{c}_{l-1}$, we use the posterior densities $p(\boldsymbol{x}_{l-1}|\mathbf{c}_{l-1}, \boldsymbol{y}_{l-1})$ and $p(\boldsymbol{S}_{l-1}|\mathbf{c}_{l-1}, \boldsymbol{y}_{l-1})$ to compute the prediction stage (24) as described in Section IV-B1. For each of these paths, there are N possible clean speech prior distributions corresponding to each c_l, creating N^2 possible paths $\{\mathbf{c}_{l-1}, c_l\}$ for which the likelihood of the observation (23) is computed. Only the best path arriving at each c_l is kept, so that

$$\forall c_l,\ \hat{\mathbf{c}}_l = \underset{\{\mathbf{c}_{l-1}, c_l\}}{\arg\max}\ p\left(\boldsymbol{y}_{1:l}, \{\mathbf{c}_{l-1}, c_l\}\right). \tag{25}$$

For each of the N retained paths, the posterior densities of $\boldsymbol{x}_l$ and $\boldsymbol{S}_l$ are computed as described in Section IV-B2.

B. Posterior Densities Computation

1) Model Prediction Step: The "Prediction" block of Fig. 2 computes the path-dependent Gaussian prior densities $p(\boldsymbol{x}_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{l-1})$. We define $\boldsymbol{F}_{l-1}$ to be the Jacobian matrix of f from the prediction equation (19) evaluated at $\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}}$ and $\boldsymbol{\mu}_{\boldsymbol{S}_{l-1}}$. It can be written as

$$\boldsymbol{F}_{l-1} = \begin{pmatrix} \boldsymbol{F}_{\boldsymbol{x}_{l-1}} & \boldsymbol{F}_{\boldsymbol{S}_{l-1}} \end{pmatrix} \tag{26}$$

with $\boldsymbol{F}_{\boldsymbol{x}_{l-1}} = \left.\frac{\partial f}{\partial \boldsymbol{x}_{l-1}}\right|_{\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}}}$ and $\boldsymbol{F}_{\boldsymbol{S}_{l-1}} = \left.\frac{\partial f}{\partial \boldsymbol{S}_{l-1}}\right|_{\boldsymbol{\mu}_{\boldsymbol{S}_{l-1}}}$.

Let us now define the augmented state

$$\boldsymbol{x}^\star_{l-1} = (\boldsymbol{x}_{l-1}, \boldsymbol{S}_{l-1})^T = \boldsymbol{\mu}_{\boldsymbol{x}^\star_{l-1}} + \delta\boldsymbol{x}^\star_{l-1} \tag{27}$$

with $\boldsymbol{\mu}_{\boldsymbol{x}^\star_{l-1}} = (\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}}, \boldsymbol{\mu}_{\boldsymbol{S}_{l-1}})^T$ and $\delta\boldsymbol{x}^\star_{l-1} \sim \mathcal{N}(0, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}} & 0 \\ 0 & \boldsymbol{\Sigma}_{\boldsymbol{S}_{l-1}} \end{pmatrix})$.

Keeping only the first two terms from the Taylor series for f [40] gives the following linear approximation:

$$f(\boldsymbol{x}^\star_{l-1}) \triangleq f(\boldsymbol{x}_{l-1}, \boldsymbol{S}_{l-1}) \approx f(\boldsymbol{\mu}_{\boldsymbol{x}^\star_{l-1}}) + \boldsymbol{F}_{l-1}\delta\boldsymbol{x}^\star_{l-1} \tag{28}$$

Computing the expected value gives us :

$$\begin{aligned} \mathrm{E}\left[f(\boldsymbol{x}^\star_{l-1})\right] &\approx \mathrm{E}\left[f(\boldsymbol{\mu}_{\boldsymbol{x}^\star_{l-1}}) + \boldsymbol{F}_{l-1}\delta\boldsymbol{x}^\star_{l-1}\right] \\ &= f(\boldsymbol{\mu}_{\boldsymbol{x}^\star_{l-1}}), \end{aligned} \tag{29}$$

which in turn gives the following covariance matrix:

$$\begin{aligned}
&\mathrm{E}\left[\left(f(\boldsymbol{x}_{l-1}^{\star})-\mathrm{E}\left[f(\boldsymbol{x}_{l-1}^{\star})\right]\right)\left(f(\boldsymbol{x}_{l-1}^{\star})-\mathrm{E}\left[f(\boldsymbol{x}_{l-1}^{\star})\right]\right)^T\right]\\
&\approx \mathrm{E}\left[\left(f(\boldsymbol{x}_{l-1}^{\star})-f(\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}^{\star}})\right)\left(f(\boldsymbol{x}_{l-1}^{\star})-f(\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}^{\star}})\right)^T\right]\\
&\approx \mathrm{E}\left[\left(\boldsymbol{F}_{l-1}\delta\boldsymbol{x}_{l-1}^{\star}\right)\left(\boldsymbol{F}_{l-1}\delta\boldsymbol{x}_{l-1}^{\star}\right)^T\right]\\
&= \boldsymbol{F}_{l-1}\mathrm{E}\left[\delta\boldsymbol{x}_{l-1}^{\star}\delta\boldsymbol{x}_{l-1}^{\star\,T}\right]\boldsymbol{F}_{l-1}^T\\
&= \boldsymbol{F}_{\boldsymbol{x}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}}\boldsymbol{F}_{\boldsymbol{x}_{l-1}}^T+\boldsymbol{F}_{\boldsymbol{S}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{S}_{l-1}}\boldsymbol{F}_{\boldsymbol{S}_{l-1}}^T. \qquad (30)
\end{aligned}$$

If we now introduce $\tilde{f}(\boldsymbol{x}_{l-1}^{\star})=\left(\boldsymbol{x}_{l-1},f(\boldsymbol{x}_{l-1}^{\star})\right)^T$, we have

$$\mathrm{E}\left[\tilde{f}(\boldsymbol{x}_{l-1}^{\star})\right]\approx\begin{pmatrix}\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}}\\ f(\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}^{\star}})\end{pmatrix} \qquad (31)$$

$$\begin{aligned}
\mathrm{Cov}\left[\tilde{f}(\boldsymbol{x}_{l-1}^{\star})\right]&\approx\begin{pmatrix}I_{2K+1} & O_{(2K+1,K)}\\ \boldsymbol{F}_{\boldsymbol{x}_{l-1}} & \boldsymbol{F}_{\boldsymbol{S}_{l-1}}\end{pmatrix}\\
&\times\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}} & 0\\ 0 & \boldsymbol{\Sigma}_{\boldsymbol{S}_{l-1}}\end{pmatrix}\begin{pmatrix}I_{2K+1} & O_{(2K+1,K)}\\ \boldsymbol{F}_{\boldsymbol{x}_{l-1}} & \boldsymbol{F}_{\boldsymbol{S}_{l-1}}\end{pmatrix}^T\\
&=\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}} & \boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}}\boldsymbol{F}_{\boldsymbol{x}_{l-1}}^T\\ \boldsymbol{F}_{\boldsymbol{x}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}} & \boldsymbol{F}_{\boldsymbol{x}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}}\boldsymbol{F}_{\boldsymbol{x}_{l-1}}^T+\boldsymbol{F}_{\boldsymbol{S}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{S}_{l-1}}\boldsymbol{F}_{\boldsymbol{S}_{l-1}}^T\end{pmatrix}. \qquad (32)
\end{aligned}$$

Therefore, by writing (19) as $\boldsymbol{x}_l=f(\boldsymbol{x}_{l-1}^{\star})+\boldsymbol{\epsilon}_l$, we can approximate the conditional joint probability of $\boldsymbol{x}_{l-1}$ and $\boldsymbol{x}_l$ by a Gaussian distribution with the following moments:

$$p(\boldsymbol{x}_{l-1},\boldsymbol{x}_l|\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})\sim\mathcal{N}\left(\boldsymbol{m}\,,\,\boldsymbol{P}\right) \qquad (33)$$

where

$$\boldsymbol{m}=\begin{pmatrix}\boldsymbol{\mu}_{x_{l-1}}\\ f(\boldsymbol{\mu}_{x_{l-1}},\boldsymbol{\mu}_{\boldsymbol{S}_{l-1}})\end{pmatrix} \qquad (34)$$

$$\boldsymbol{P}=\begin{pmatrix}\boldsymbol{\Sigma}_{x_{l-1}} & \boldsymbol{\Sigma}_{x_{l-1}}\boldsymbol{F}_{x_{l-1}}^T\\ \boldsymbol{F}_{x_{l-1}}\boldsymbol{\Sigma}_{x_{l-1}} & \boldsymbol{F}_{x_{l-1}}\boldsymbol{\Sigma}_{x_{l-1}}\boldsymbol{F}_{x_{l-1}}^T+\boldsymbol{F}_{\boldsymbol{S}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{S}_{l-1}}\boldsymbol{F}_{\boldsymbol{S}_{l-1}}^T+\boldsymbol{Q}_l\end{pmatrix} \qquad (35)$$

where the means and covariance matrices of $\boldsymbol{x}_{l-1}$ and $\boldsymbol{S}_{l-1}$ are the moments of the posterior distributions $p(\boldsymbol{x}_{l-1}|\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})$ and $p(\boldsymbol{S}_{l-1}|\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})$. We therefore have for the marginal probability density of $\boldsymbol{x}_l$:

$$p(\boldsymbol{x}_l|\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})\sim\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}},\,\boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}\right) \qquad (36)$$

with

$$\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}=f(\boldsymbol{\mu}_{\boldsymbol{x}_{l-1}},\boldsymbol{\mu}_{\boldsymbol{S}_{l-1}}) \qquad (37)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}=\boldsymbol{F}_{\boldsymbol{x}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{x}_{l-1}}\boldsymbol{F}_{\boldsymbol{x}_{l-1}}^T+\boldsymbol{F}_{\boldsymbol{S}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{S}_{l-1}}\boldsymbol{F}_{\boldsymbol{S}_{l-1}}^T+\boldsymbol{Q}_l \qquad (38)$$

giving us the solution to (24).

2) Observation Update Step: This section describes the "Clean Speech HMM" and "Likelihood Computation" blocks of Fig. 2. These compute the likelihood of the observation $p(\boldsymbol{y}_l|c_l,\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})$ for each of the N^2 possible paths $\{\mathbf{c}_{l-1},c_l\}$ as well as the posterior densities of the state vector and clean speech log-power, $p(\boldsymbol{x}_l|c_l,\mathbf{c}_{l-1},\boldsymbol{y}_l,\boldsymbol{y}_{l-1})$ and $p(\boldsymbol{S}_l|c_l,\mathbf{c}_{l-1},\boldsymbol{y}_l,\boldsymbol{y}_{l-1})$.

The assumption in (9) that speech, reverberation and noise powers add to form the observed power imposes a nonlinear constraint in the log-power domain. Similar to the derivations in Section IV-B1, we can use a first order Taylor series approximation of h in the observation equation (20) to obtain mean and covariance for the approximately Gaussian joint distribution of $\boldsymbol{y}_l$ and $\boldsymbol{x}_l$. We define $\boldsymbol{H}_l$ as the Jacobian matrix of $h(\boldsymbol{x}_l,\boldsymbol{S}_l)$ evaluated at $(\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}},\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}})$ so that

$$\boldsymbol{H}_l=\left(\boldsymbol{H}_{\boldsymbol{x}_l}\;\boldsymbol{H}_{\boldsymbol{S}_l}\right). \qquad (39)$$

The mean, $\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}$, and covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}$, of the predicted pdf of $\boldsymbol{x}_l$ for the path originating at $\mathbf{c}_{l-1}$ are given in (37),(38). The mean, $\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}}$, and covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}$, of the prior pdf associated with state c_l are learned during training.

Using similar derivations to (28)-(32), it follows that for the path defined by $\{\mathbf{c}_{l-1},c_l\}$ we have:

$$p(\boldsymbol{x}_l,\boldsymbol{y}_l|c_l,\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})\sim\mathcal{N}\left(\boldsymbol{m}_{\boldsymbol{xy}}\,,\,\boldsymbol{C}_{\boldsymbol{xy}}\right) \qquad (40)$$

where

$$\boldsymbol{m}_{\boldsymbol{xy}}=\begin{pmatrix}\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}\\ h(\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}},\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}})\end{pmatrix} \qquad (41)$$

$$\begin{aligned}
\boldsymbol{C}_{\boldsymbol{xy}}&=\begin{pmatrix}I_{(K,2K+1)} & O_K\\ \boldsymbol{H}_{\boldsymbol{x}_l} & \boldsymbol{H}_{\boldsymbol{S}_l}\end{pmatrix}\times\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} & 0\\ 0 & \boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}\end{pmatrix}\\
&\times\begin{pmatrix}I_{(K,2K+1)} & O_K\\ \boldsymbol{H}_{\boldsymbol{x}_l} & \boldsymbol{H}_{\boldsymbol{S}_l}\end{pmatrix}^T+\begin{pmatrix}0 & 0\\ 0 & \boldsymbol{M}_l\end{pmatrix}\\
&=\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} & \boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}\boldsymbol{H}_{\boldsymbol{x}_l}^T\\ \boldsymbol{H}_{x_l}\boldsymbol{\Sigma}_{x_l|c_{l-1}} & \boldsymbol{H}_{x_l}\boldsymbol{\Sigma}_{x_l|c_{l-1}}\boldsymbol{H}_{x_l}^T+\boldsymbol{H}_{\boldsymbol{S}_l}\boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}\boldsymbol{H}_{\boldsymbol{S}_l}^T+\boldsymbol{M}_l\end{pmatrix}. \qquad (42)
\end{aligned}$$

The observation noise covariance matrix, $\boldsymbol{M}_l$, in (42) represents the uncertainty between the model of (13) and the actual observations. It is the sum of a fixed component that is a function of the filterbank parameters $b_{k,\tilde{k}}$ in (3) and another that depends on the estimated mean and variance of the observation, $\breve{Y}(l,k)$. Detailed expressions for these two components are given in Appendices A and B respectively.

We therefore have the likelihood of the observation

$$p(\boldsymbol{y}_l|c_l,\mathbf{c}_{l-1},\boldsymbol{y}_{l-1})\sim\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{y}_l},\boldsymbol{\Sigma}_{\boldsymbol{y}_l}\right) \qquad (43)$$

with

$$\boldsymbol{\mu}_{\boldsymbol{y}_l}=h(\boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}},\boldsymbol{\mu}_{\boldsymbol{S}_{c_l}}) \qquad (44)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_l}=\boldsymbol{H}_{\boldsymbol{x}_l}\boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}}\boldsymbol{H}_{\boldsymbol{x}_l}^T+\boldsymbol{H}_{\boldsymbol{S}_l}\boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}\boldsymbol{H}_{\boldsymbol{S}_l}^T+\boldsymbol{M}_l \qquad (45)$$

and the posterior pdf of $\boldsymbol{x}_l$ [40], [41]

$$p(\boldsymbol{x}_l|c_l,\mathbf{c}_{l-1},\boldsymbol{y}_l,\boldsymbol{y}_{l-1})\sim\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{x}_l},\boldsymbol{\Sigma}_{\boldsymbol{x}_l}\right) \qquad (46)$$

with

$$\boldsymbol{\mu}_{\boldsymbol{x}_l} = \boldsymbol{\mu}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} + \boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} \boldsymbol{H}_{\boldsymbol{x}_l}^T \boldsymbol{\Sigma}_{\boldsymbol{y}_l}^{-1} [\boldsymbol{y}_l - \boldsymbol{\mu}_{\boldsymbol{y}_l}] \tag{47}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_l} = \boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} - \boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} \boldsymbol{H}_{\boldsymbol{x}_l}^T \boldsymbol{\Sigma}_{\boldsymbol{y}_l}^{-1} \boldsymbol{H}_{\boldsymbol{x}_l} \boldsymbol{\Sigma}_{\boldsymbol{x}_l|\mathbf{c}_{l-1}} \tag{48}$$

which uses a similar approach to the implementation of an Extended Kalman Filter (EKF). Equations (43)-(45) can then be used to compute the joint likelihood of the observations and sequence of states in (22).

Using a similar method to (40)-(42), we can approximate the joint distribution of the observation and clean speech log-power as a Gaussian distribution to obtain

$$p(\boldsymbol{S}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_l, \boldsymbol{y}_{l-1}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{S}_l}, \boldsymbol{\Sigma}_{\boldsymbol{S}_l}) \tag{49}$$

with

$$\boldsymbol{\mu}_{\boldsymbol{S}_l} = \boldsymbol{\mu}_{\boldsymbol{S}_{c_l}} + \boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}} \boldsymbol{H}_{\boldsymbol{S}_l}^T \boldsymbol{\Sigma}_{\boldsymbol{y}_l}^{-1} [\boldsymbol{y}_l - \boldsymbol{\mu}_{\boldsymbol{y}_l}] \tag{50}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{S}_l} = \boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}} - \boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}} \boldsymbol{H}_{\boldsymbol{S}_l}^T \boldsymbol{\Sigma}_{\boldsymbol{y}_l}^{-1} \boldsymbol{H}_{\boldsymbol{S}_l} \boldsymbol{\Sigma}_{\boldsymbol{S}_{c_l}}. \tag{51}$$

The N-best paths can then be pruned according to (25), and the associated posterior densities can then be used in order to update the reverberation parameters estimate $\boldsymbol{\pi}_l$ and compute the gain $\check{\boldsymbol{W}}_l$.

Numerical errors can arise when computing $\boldsymbol{\Sigma}_{\boldsymbol{y}_l}$ using (45) that may lead to the estimated covariance matrix being non-positive definite and preventing the computation of the likelihood of the observation. This can especially happen when the observation noise is very low. Though not described in detail here, this problem can be solved by implementing the Square Root version of the Extended Kalman Filter-type update (SR-EKF). By factorizing $\boldsymbol{\Sigma}_{\boldsymbol{x}_l}$ and $\boldsymbol{\Sigma}_{\boldsymbol{S}_l}$ in a $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ form where $\boldsymbol{U}$ is a unit upper triangular matrix and $\boldsymbol{D}$ is a diagonal matrix, we can carry the updates on both these matrices and ensure that the covariance matrices of $p(\boldsymbol{y}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_{l-1})$, $p(\boldsymbol{x}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_l, \boldsymbol{y}_{l-1})$ and $p(\boldsymbol{S}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_l, \boldsymbol{y}_{l-1})$ remain positive-definite. This is achieved by using the Bierman-Thornton SR-EKF, which is a combination of the Square-Root implementations proposed in [42], [43].

3) On the Approximation of Transformed Distributions: In this section we look at how well the Taylor series approximation of h allows us to approximate the transformed pdfs. To do so, for clarity we consider the 2-dimensional case with random variables A and B, in which we assume no observation noise is present. We assume that A and B are jointly Gaussian as in Fig. 3 (a) where the log-probability density values have been scaled to match the displayed colormap. On the plot, the mean is marked by a cross and the unit standard deviation contour by an ellipse. The dotted line indicates the constraint $\log(\exp(A) + \exp(B)) = 0$ analogous to (9). We can approximate the constrained distribution (i.e. the posterior distribution) by computing the empirical mean and covariance of the points lying on the contour $\log(\exp(A) + \exp(B)) = 0$. The Gaussian distribution with the empirical mean and covariance is shown in Fig. 3 (b).

The constrained distribution computed using a first order Taylor series approximation of the nonlinear constraint is shown in Fig. 3 (c). There is a large underestimate of the variance in the direction orthogonal to the tangent of the non-linear constraint. This can be explained by the first order linearization of the constraint, which forces the constrained distribution to lie on the tangent. If the original unconstrained distribution is very close to one of the extremes of the constraint, corresponding to a highly positive or highly negative SNR, this approximation is accurate. However, the approximated covariance is too small at the maximum curvature point of the constraint.

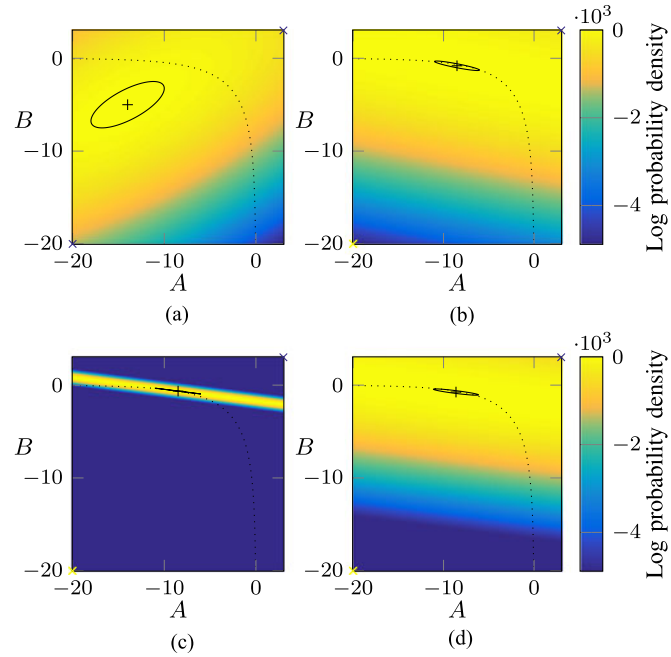


Fig. 3. Two-dimensional case : A and B are jointly Gaussian distributed. Unconstrained prior (a), empirically computed constrained posterior (b) distributions. Using Taylor series approximation of the nonlinear constraint, the first-order (c) and second-order (d) approximation of the constrained distribution are shown.

Although not used in our implementation, these approximation inaccuracies can be reduced by using a second-order Taylor series approximation of our constraint which gives the approximated constrained distribution shown in Fig. 3 (d). The result is closer to the empirically computed distribution, suggesting that better results could be achieved using a second-order Taylor series approximation in Section IV-B2. This adds an additional term to the covariance matrix of the marginal distribution of the observation, of the form

$$\sum_{i,j} \boldsymbol{e}_i \boldsymbol{e}_j^T \operatorname{tr}\left[\boldsymbol{H}_{xs}^{(i)} \boldsymbol{\Sigma}_{x_l} \boldsymbol{H}_{xs}^{(j)} \boldsymbol{\Sigma}_{x_l}\right] \tag{52}$$

with $\boldsymbol{H}_{xs}^{(i)}$ the Hessian of h at output dimension i, tr[.] indicating the trace of the matrix, and $\boldsymbol{e}_i = [0, 0, \ldots, 1, \ldots, 0, 0]^T$ where the 1 is at position i. As this requires substantial additional computation, we instead use a first-order approximation with an additional observation noise term compensating for the underestimated covariance while remaining computationally efficient. A detailed derivation of this additional noise term is given in Appendix B.

C. Reverberation Parameters Estimation

In Sections IV-A & IV-B, the reverberation parameters $\boldsymbol{\alpha}$ and $\boldsymbol{d}$ are assumed fixed in order to compute the moments of the probability distributions in the prediction step. However, as we do not assume a perfect initialization for these parameters, and as the DRR can change dynamically due to movement of the speaker or changes in the acoustic environment, we need to update our reverberation parameters estimates adaptively.

We define

$$\boldsymbol{\pi}_l = \left(\log\left(\frac{\boldsymbol{\alpha}_l}{1-\boldsymbol{\alpha}_l}\right), \log\left(\frac{\boldsymbol{d}_l}{1-\boldsymbol{d}_l}\right)\right)^T \tag{53}$$

to be the vector of transformed reverberation parameters, where we map the range $(0,1)$ to $(-\infty,+\infty)$ to avoid the need for range constraints on the elements of $\boldsymbol{\pi}_l$. In the following, we identify global random variables that take into account all paths in the HMM with an overbar, $^{-}$.

We define the following dynamic equations describing the evolution of the reverberation parameters:

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_{l-1} + \boldsymbol{\omega}_l \tag{54}$$

$$\bar{\boldsymbol{\mu}}_{\boldsymbol{R}_l} = g(\boldsymbol{\pi}_l, \bar{\boldsymbol{\mu}}_{\boldsymbol{R}_{l-1}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{S}_{l-1}}) + \boldsymbol{\psi}_l \tag{55}$$

where $\bar{\boldsymbol{\mu}}_{\boldsymbol{R}_l}$, the mean of the global posterior density of $\boldsymbol{R}_l$, acts as observation, $\bar{\boldsymbol{\mu}}_{\boldsymbol{R}_{l-1}}$ and $\bar{\boldsymbol{\mu}}_{\boldsymbol{S}_{l-1}}$ act as fixed system parameters, $\boldsymbol{\omega}_l \sim \mathcal{N}(0, \boldsymbol{U}_l)$ and $\boldsymbol{\psi}_l \sim \mathcal{N}(0, \boldsymbol{V}_l)$. $\boldsymbol{U}_l$ controls how much the reverberation parameters are allowed to change from one frame to the next, while $\boldsymbol{V}_l$ represents errors in the model of (12), of which g is a direct implementation.

Assuming we have for each of the N paths $\mathbf{c}_l$ the posterior pdfs of $\boldsymbol{x}_l$ and $\boldsymbol{S}_l$, we can compute the global posterior densities as

$$\begin{aligned} &p(\bar{\boldsymbol{x}}_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{1:l}) \\ &\quad = \sum_{c_l} p(c_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{1:l}) p(\boldsymbol{x}_l|c_l, \mathbf{c}_{l-1}, \boldsymbol{y}_l, \boldsymbol{y}_{l-1}) \end{aligned} \tag{56}$$

with the normalized path probabilities defined as

$$p(c_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{1:l}) = \frac{p(c_l, \boldsymbol{y}_l, \mathbf{c}_{l-1}, \boldsymbol{y}_{1:l-1})}{\sum_{c_l} p(c_l, \boldsymbol{y}_l, \mathbf{c}_{l-1}, \boldsymbol{y}_{1:l-1})} \tag{57}$$

and similarly for $p(\bar{\boldsymbol{S}}_l|\mathbf{c}_{l-1}, \boldsymbol{y}_{1:l})$. The means of these global pdfs are then directly calculated as the weighted sum of the means of each individual path mixture. The mean of the global posterior distribution of the reverberation log-power, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{R}_l}$, is directly extracted from that of $\boldsymbol{x}_l$.

From (54)-(55) we can therefore obtain the first and second-order moments of the posterior distribution for $\boldsymbol{\pi}_l$ using:

$$\boldsymbol{\mu}_{\boldsymbol{\pi}_l} = \boldsymbol{\mu}_{\boldsymbol{\pi}_{l-1}} + \boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l|l-1}} \boldsymbol{G}^T \boldsymbol{C}_{\boldsymbol{\pi}}^{-1} \boldsymbol{e}_{\bar{\boldsymbol{R}}_l} \tag{58}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\pi}_l} = \boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l|l-1}} - \boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l|l-1}} \boldsymbol{G}^T \boldsymbol{C}_{\boldsymbol{\pi}}^{-1} \boldsymbol{G} \boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l|l-1}} \tag{59}$$

where $\boldsymbol{e}_{\bar{\boldsymbol{R}}_l} = \bar{\boldsymbol{\mu}}_{\boldsymbol{R}_l} - g(\boldsymbol{\mu}_{\boldsymbol{\pi}_{l-1}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{R}_{l-1}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{S}_{l-1}})$ is the error in the predicted mean reverberation power, $\boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l|l-1}} = (\boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l-1}} + \boldsymbol{U}_l)$ is the covariance matrix of the predicted $\boldsymbol{\pi}_l$ of (54), $\boldsymbol{G} = \frac{\partial g}{\partial \boldsymbol{\pi}_l}\Big|_{\boldsymbol{\mu}_{\boldsymbol{\pi}_{l-1}}}$ is the Jacobian matrix of g and $\boldsymbol{C}_{\boldsymbol{\pi}} = \boldsymbol{G}\boldsymbol{\Sigma}_{\boldsymbol{\pi}_{l|l-1}}\boldsymbol{G}^T + \boldsymbol{V}_l$.

The resulting algorithm is therefore a two-stage approach. First we fix the reverberation parameters in order to compute the likelihood of each path in the HMM, so as to get the posterior probability densities of $\boldsymbol{x}_l$ and $\boldsymbol{S}_l$ for the best path arriving at each possible state in the HMM. Then, the means of the global posterior densities are computed and fixed in order to update the reverberation parameters using (58)-(59).

V. Performance Evaluation

The evaluation of the proposed algorithm on actual reverberant noisy data is divided into two parts. First, because most objective metrics for speech quality and intelligibility are intrusive, we generate simulated reverberant data by convolving anechoic speech with measured room impulse responses so that we can have access to the original clean speech. Second, the algorithm is tested on real data, i.e. actual reverberant and noisy recordings for which no target clean signal is available.

We compare our method with the single-channel scheme of Cauchi *et al*. [14] as it was the only single-channel method participating in the REVERB Challenge [5] which managed to reduce the perceived amount of reverberation appreciably while significantly improving the overall speech quality [21]. We therefore consider this competing method to be state-of-the-art. The parameters of this competing method correspond to those described in [14] and the implementation was generously provided by the author. A difference between the two algorithms is that [14], although a spectral enhancement algorithm suitable for real-time implementation, requires an external estimate of the broadband T_{60} which is obtained using the utterance-based algorithm presented in [44]. The proposed method, in contrast, does not require prior knowledge of the reverberation parameters and is implemented in an online manner computing the spectral gain at each time frame. On a laptop equipped with an Intel Core i5 processor, the average real-time factors of the two methods were measured to be 0.17 for [14] and 3.65 for the proposed method. An implementation in MATLAB of the proposed method is available as `spendred.m` in the VOICEBOX toolbox [32].

A. Implementation Details

1) HMM States Learning: To train the mean and covariance matrices of each state in the HMM, we use a purely data-driven technique, which gives us the ability to work with any clean speech dataset with the minimum amount of adaptation effort. This can also help to make the set of states less language-dependent and to provide better generalization. To determine a representative set of states, we used the k-means [45], [46] feature-learning technique, as it remains a method of choice in many practical scenarios thanks to its scalability [47]. Viewing the k-means algorithm from a Bayesian perspective, minimizing the Euclidean distance is equivalent to maximizing the likelihood of the clusters according to Gaussian distributions with identity covariance matrices. This fits well with the assumptions of our model, and we can perform the clustering directly on the Mel-frequency log-spectral powers. We used the k-means implementation available in [32], and computed 15 separate

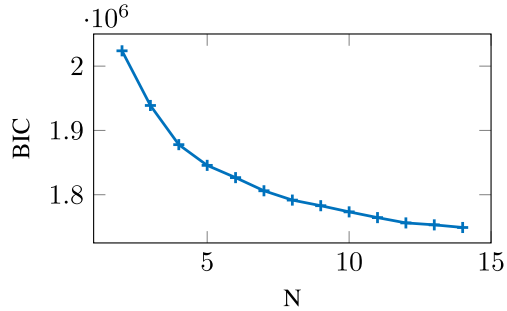


Fig. 4. Bayesian Information Criterion (BIC) computed for different values of N the number of clusters used in the k-means algorithm.

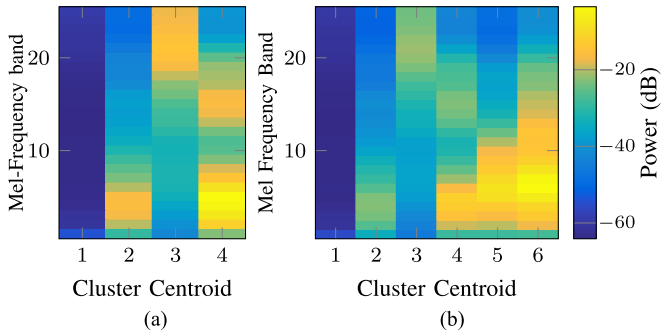


Fig. 5. Means of the log-power clean speech HMM states obtained through k-means with (a) 4 clusters and (b) 6 clusters.

instances with random initialization for N, the number of clusters, varying from 2 to 14. Such a low number of states may seem surprising, as a much higher number of dictionary elements has been reported to be necessary in speech enhancement applications using NMF-based techniques [48]. Here, however, we look at log-power spectral frames on a Mel-frequency scale having broad frequency bands and the learned states are used only to provide prior probabilities in a Bayesian inference context rather than used directly in a Wiener filter as in [28], [30]. We used the training set of the TIMIT database [49], normalized the input speech signals to 0 dB active level [32], [33], obtained STFT frames of 30 ms with 5 ms frame increment and computed the log-power in each Mel-frequency band for each frame.

The Bayesian Information Criterion (BIC) [50] was computed for each value of N according to

$$\mathrm{BIC} = -2\log(\mathcal{L}) + \log(n)KN \tag{60}$$

and is plotted in Fig. 4. $\mathcal{L}$ is the likelihood of the observed data and n is the number of data points in the observed data. From a clustering point of view, the BIC gives an idea of how well the clusters can explain the whole dataset. It appears from Fig. 4 that the BIC does not improve significantly for 10 clusters or more. However, from an inference point of view, the HMM states are only used as possible prior density functions for the clean speech, reducing even further the need for a set of states able to perfectly represent any clean speech signal directly. This allows us to use a low number of states, and in our experiments we have chosen $N = 4$ or $N = 6$. The state means obtained for $N = 4$ are shown in Fig. 5 (a); these states correspond approximately to a silence state, a voiced state, an unvoiced state, and a voiced/unvoiced combination. The state means obtained for $N = 6$ are shown in Fig. 5 (b); the first four are similar to those of Fig. 5 (a) while the remaining two correspond to additional voiced spectra. The results for simulated data are presented for an implementation with 4 states in Section V-C, while both 4 and 6-state implementations are used to evaluate the performance on live recordings in Section V-D.

2) Algorithm Parameters: In order to obtain better dereverberation and denoising performance, we used $\beta = 1$, i.e. a Wiener gain, $\eta = 2$ and $\lambda_s = 0.95$ in (18). We have found that the proposed algorithm is not very sensitive to the initial values used for the reverberation parameters and the same initial values were used for all reported experiments. The initial values for the frequency-dependent $\boldsymbol{\alpha}$ were chosen to correspond to the subband T_{60} values averaged over all RIRs in [51] and [52]. We initialized $\boldsymbol{d}$ to correspond to linearly spaced subband DRR values ranging from -2 dB in the lowest Mel-frequency band to 8 dB in the highest band according to (8). The first 100 ms of each recording were assumed to be noise and were used to initialize the mean and covariance of the noise log-power in $\boldsymbol{x}_0$. Reverberation log-power was initialized at 10 dB below the noise and the clean speech global gain was initialized to -5 dB. The STFT analysis used 30 ms Hann-windowed frames with a frame increment of 5 ms. The number of Mel-frequency bands was set to $K = 25$.

B. Evaluation Metrics

Six different metrics were used in order to evaluate the algorithms: the Cepstrum Distance (CD) [53], the Frequency-weighted Segmental SNR (FWSegSNR) [54], the Reverberation Decay Tail ($\mathrm{R_{DT}}$) [55], the normalized Speech-to-Reverberation Modulation energy Ratio ($\mathrm{SRMR_{norm}}$) [56] (available at [57]), the Short-Time Objective Intelligibility score (STOI) [58] (available at [59]) and the Perceptual Evaluation of Speech Quality (PESQ) [60]. The STOI scores were mapped to a percentage of words correctly recognized using the mapping function provided in [58] in order to make results easier to read and interpret. The implementations of CD and FWSegSNR were taken from [5], while we used a direct implementation of [55] for $\mathrm{R_{DT}}$.

CD has been reported to be well correlated with the overall quality of processed noisy speech as well as the perceived level of reverberation [21], [61], [62]. However, conflicting results have been found regarding its correlation with the overall quality of enhanced reverberant speech [21], [62], and it has been found to correlate poorly with speech intelligibility [63]. $\mathrm{R_{DT}}$ and $\mathrm{SRMR_{norm}}$ have been found to correlate well with the perceived level of reverberation [62], [64]. The FWSegSNR and PESQ measures have generally been reported to correlate well with overall quality and intelligibility [58], [61]–[63]. Finally, STOI has been found to be highly correlated with intelligibility for time-frequency weighted noisy speech [58].

C. Simulated Data

In order to test the performance of our algorithm in challenging scenarios, we use the Acoustic Characterisation of Environments (ACE) Challenge Corpus [65], which was developed to

TABLE I
TABLE DETAILING INFORMATION ABOUT RIRS FROM THE ACE CORPUS USED TO CREATE THE SIMULATED DATA

Acoustic Condition	Room	Config.	T_{60} (s)	DRR (dB)
A	Lobby	1	0.81	6.47
B	Lobby	2	0.77	3.25
C	Lecture	1	1.33	8.94
D	Lecture	2	1.29	4.96
E	Meeting	1	0.38	5.00
F	Meeting	2	0.38	8.38
G	Office	1	0.40	2.44
H	Office	2	0.40	−2.27

evaluate algorithms for blind estimation of acoustic parameters in the presence of noise. The corpus provides multi-channel RIRs as well as noises recorded in-situ for various acoustic spaces (lecture rooms, offices, meeting rooms, lobby). For the single channel case, measured impulse responses and corresponding noises are provided for two different source-receiver positions within each room. There are three noise types: fan noise, ambient noise and babble noise. All noise types were recorded in situ using identical microphone configurations and are therefore consistent with the measured impulse responses. The babble noise was recorded using actual talkers in each room, and the RIRs were measured with the talkers still present inside the room.

From the ACE challenge clean speech corpus, we selected sound files from 14 speakers in total (5 females and 9 males), each uttering a free-speech sentence approximately 10 seconds long describing where they live. The anechoic speech files were convolved with one of 8 RIRs corresponding to 4 different rooms and 2 source-microphone positions within each room. Table I gives the broadband T_{60} and DRR values measured from the impulse responses using [66] and [11].

For each measured impulse response, the corresponding ambient, fan and babble noises were used and random portions of these recordings were added to the reverberant speech at 0, 10 and 20 dB SNR. This makes a total of 1008 noisy and reverberant speech files.

First, in order to assess the dereverberation performance of our algorithm, we show in Fig. 6 the average score for each metric in the case of 20 dB SNR only, averaging the results over the three noise types. An SNR of 20 dB is still a realistic environment, but the noise has a limited degradation effect and therefore we expect the results to be dominated by the dereverberation performance of the methods.

The proposed method leads to the lowest Cepstral Distance (plot a), highest Frequency-weighted Segmental SNR (plot b) and lowest reverberation decay tail (plot c) for all acoustic conditions, suggesting the proposed method achieves better dereverberation performance than [14]. Both algorithms yield very similar PESQ (plot f) and STOI (plot e) results, with a slight improvement of predicted intelligibility in the most reverberant case (D), and a near-constant PESQ improvement of about 0.2 over unprocessed speech. This seems to suggest the proposed method improves speech quality as much as the competing

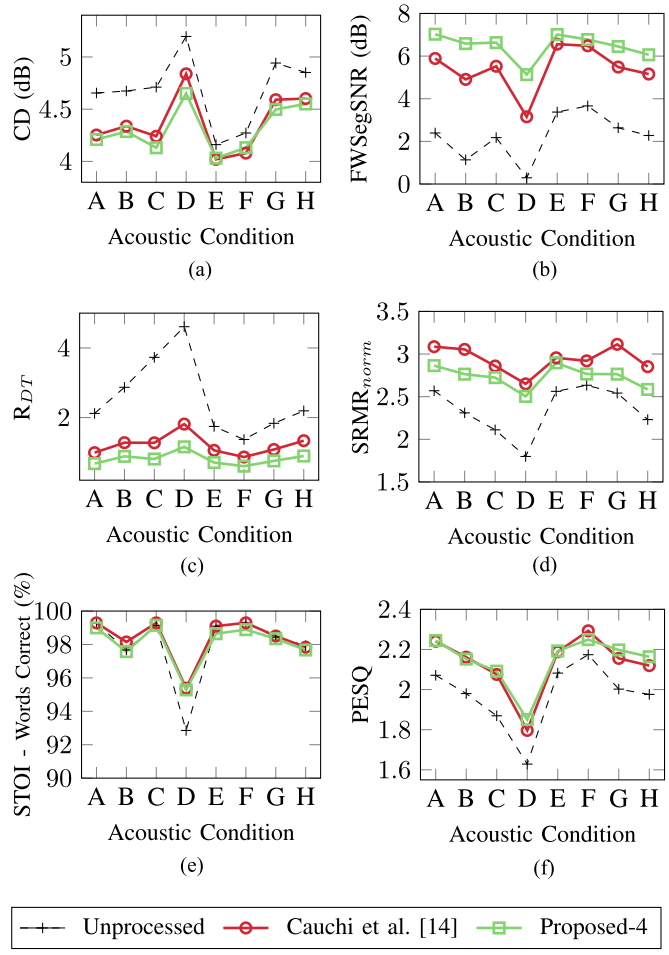


Fig. 6. Results comparing the two speech enhancement methods on simulated data (a) - Cepstrum Distance (dB), the lower the better (b) Frequency-Weighted Segmental SNR, the higher the better (c) - Reverberation Decay Tail, the lower the better (d) - Normalized version of Speech to Reverberation Modulation Energy Ratio, the higher the better (e) - STOI scores mapped to words correctly recognized in %, the higher the better (f) - PESQ scores, the higher the better.

one without degrading intelligibility. The proposed method achieves better results than unprocessed speech with respect to the $\mathrm{SRMR_{norm}}$ metric, but does less well than [14]. This contradicts the $\mathrm{R_{DT}}$ result as it suggests a higher perceived reverberation than [14], however the validity of the $\mathrm{SRMR_{norm}}$ metric for use with processed speech signals has not been studied.

In order to study the robustness of both methods to noise, we show box plots of the differential (Δ) scores obtained for each metric, separated for each SNR and each noise type and averaged across all acoustic conditions. On the box plots, the interquartile range is shown by a coloured box, the median of the distribution is shown by a horizontal line, and the mean of the distribution is shown by a circle. For each result, a 0 score indicates no change in the metric compared to unprocessed speech, and a positive result indicates a higher metric score.

Fig. 7 indicates that, apart from the babble noise case, the proposed algorithm achieves lower Cepstral Distance than [14], especially at low SNRs, indicating that it is better able to deal with heavy noise. Furthermore, as can be seen in Fig. 8, the

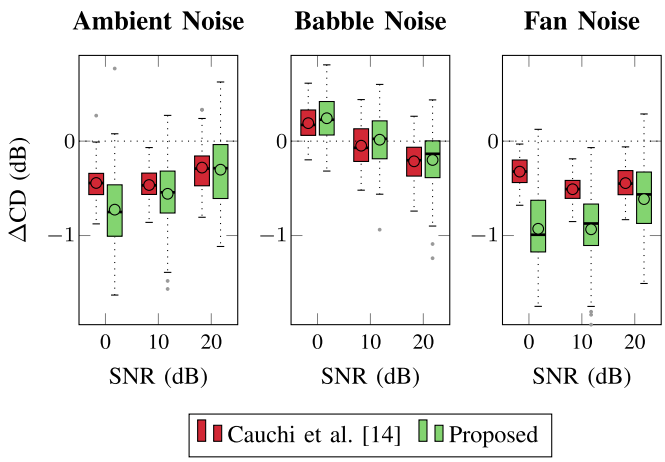


Fig. 7. Differential Cepstral Distance for different noise conditions, averaged across all acoustic scenarios.

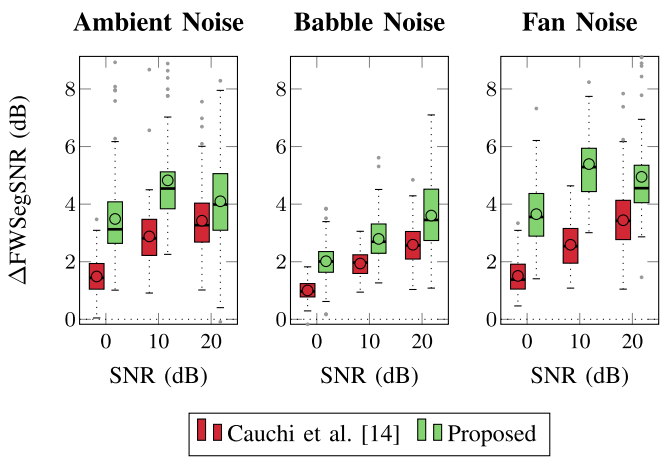


Fig. 8. Differential Frequency-Weighted Segmental SNR for different noise conditions, averaged across all acoustic scenarios.

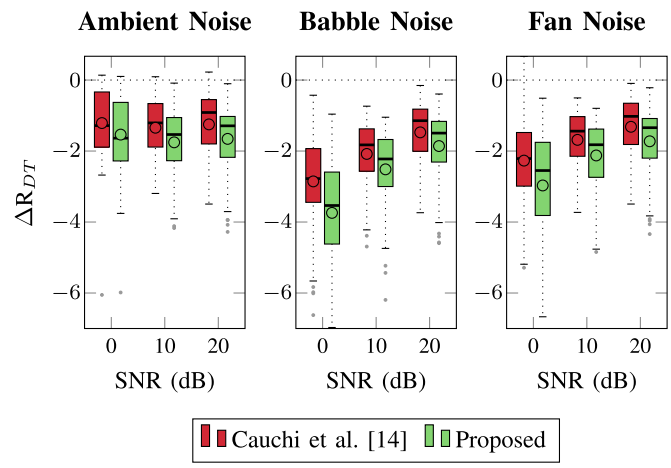


Fig. 9. Differential R_{DT} scores obtained for different noise conditions, averaged across all acoustic scenarios.

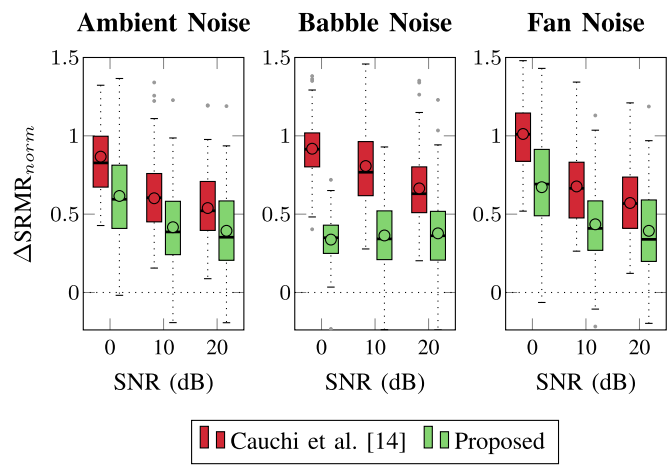


Fig. 10. Differential $SRMR_{norm}$ scores for different noise conditions, averaged across all acoustic scenarios.

higher FWSegSNR scores achieved by the proposed method in all cases seem to suggest better dereverberation as well as better noise reduction properties.

Fig. 9 shows that even when the SNR is low, the proposed algorithm achieves lower R_{DT} scores than [14]. This indicates that even in the presence of heavy noise, it is able to reduce the decay tail of the reverberation significantly. Unsurprisingly, both methods achieve very low R_{DT} scores in babble noise. Indeed, with the ACE challenge corpus the babble noise was recorded using talkers in situ, giving much more information about the acoustic properties of the whole recording. Fig. 10 confirms the earlier observation that the proposed method achieves lower $SRMR_{norm}$ scores compared to [14], although they are almost always greater than those of the unprocessed speech.

As can be seen in Fig. 11, the predicted intelligibility is slightly worse with the proposed method than with [14]. However, as was seen in Fig. 6 (e), the predicted intelligibility of the test signals was well above 90% in all cases so these small differences will have little effect. The PESQ scores, shown in Fig. 12, show a consistent improvement for both algorithms relative to unprocessed speech with the proposed method having marginally higher scores than [14].

Overall, it seems the proposed method achieves better dereverberation and denoising performance while improving speech quality and preserving speech intelligibility. It also seems that [14] deals with babble noise slightly better, which is unsurprising since our clean speech model cannot distinguish babble noise from wanted speech. Tests using 6 states in the HMM were also carried out, but the results were almost identical to those using 4 states and are not presented here.

D. Real Data

We used the real data section of the evaluation set of the REVERB Challenge [21], which corresponds to the Multi-Channel Wall Street Journal Audio Visual Corpus [67]. The data was recorded in a room using real talkers and at two different source-microphone positions, i.e. near and far. Because no reference signal is available, and in order to gain some insight into how well the dereverberation methods worked on this dataset, we used the baseline ASR systems from the REVERB challenge

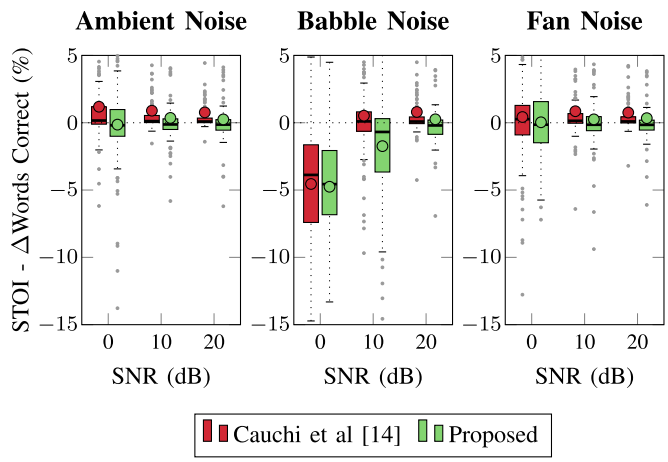


Fig. 11. Differential mapped STOI scores for different noise conditions, averaged across all acoustic scenarios.

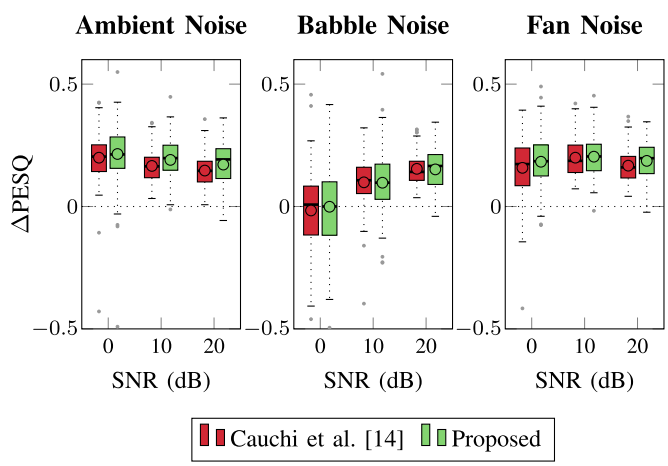


Fig. 12. Differential PESQ scores for different noise conditions, averaged across all acoustic scenarios.

to obtain WER scores. The proposed algorithm was evaluated using both 4 states and 6 states in the HMM.

All methods were tested on two baseline speech recognition engines from [21]. The baseline systems were both based on HTK, using a triphone GMM-HMM recognizer that has been trained on clean speech data only. One version of the engine used Constrained Maximum Likelihood Linear Regression (CMLLR) speaker adaptation while the other did not. Fig. 13 shows the reduction in WER achieved by [14] and by the proposed method using a 4 or 6-state HMM.

The proposed method achieves lower WER than unprocessed speech, with significantly better results obtained when using a 6-state HMM for the clean speech model, but still higher WER than the competing method. Although the audible quality of the recordings has been substantially improved, we believe that our method may introduce more artifacts detrimental to such ASR systems than [14]. Audio recordings processed by the 6-state implementation as well as the listening test results presented below are available from http://www.commsp.ee.ic.ac.uk/˜sap/sicenspeech/.

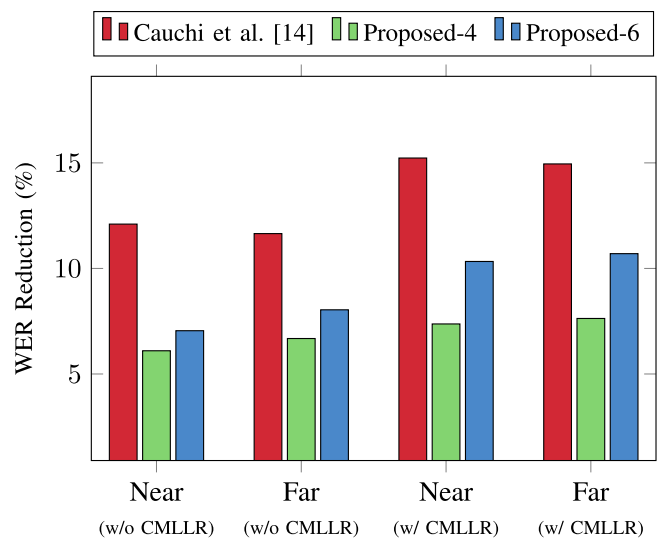


Fig. 13. Average WER reduction for the different acoustic conditions of the REVERB challenge real data.

E. Listening Test

Although objective metrics are a good indication of an algorithm's performance, it has been hypothesized that no instrumental measure can capture the subjective sense of overall speech quality [21]. Therefore a listening test similar to the multi-stimuli with hidden reference and anchor (MUSHRA) [68] test was used in order to assess the overall quality and amount of perceived reverberation before and after processing. The ambient noise level and the headphones used in the experiment were not controlled and varied between participants.

The 13 self-reported normal-hearing participants, all experts in acoustic signal processing, each performed 8 tests: 4 tests rating the perceived level of reverberation on a scale ranging from 0 (not reverberant) to 100 (very reverberant), and 4 tests rating the overall speech quality on a scale going from 0 (bad) to 100 (excellent). Post-screening was performed after the test in order to remove results where participants failed to identify the hidden reference.

For each test, the participants were asked to compare four randomly-ordered unmarked samples: (i) a hidden reference, (ii) a noisy reverberant anchor signal, (iii) the anchor signal processed by [14] and (iv) the anchor signal processed by the proposed method with a 6-state HMM. The hidden reference was a clean speech utterance convolved with an RIR from [52] with very low T_{60} (0.18 s) and high DRR (5 dB), as in MUSHRA for reverberant speech or MUSHRAR proposed in [64]. To form the anchor signals, clean speech utterances from the ACE challenge corpus [65] were first convolved with RIRs B, D, E and H from Table I to create reverberant signals with $0.38\,\mathrm{s} \leq T_{60} \leq 1.29\,\mathrm{s}$ and $-2.27\,\mathrm{dB} \leq \mathrm{DRR} \leq 5\,\mathrm{dB}$. For the tests that evaluated speech quality, these reverberant signals were then degraded by adding "ambient noise" from [65] at 0 dB or 10 dB SNR. For the tests that evaluated reverberation reduction they were degraded by adding "babble noise" from [65] at 30 dB SNR.

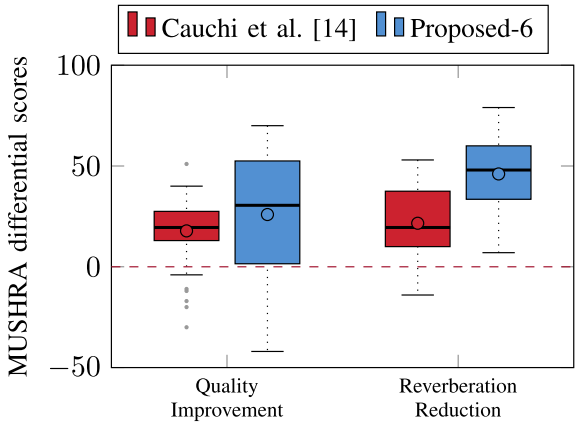


Fig. 14. Listening test results. MUSHRA differential scores corresponding to the overall speech quality improvement and perceived reverberation reduction.

The results are shown in Fig. 14 which presents differential MUSHRA scores between the unprocessed reverberant and noisy speech and the two processed versions. These differential scores can be viewed as measuring the overall quality improvement and reverberation reduction provided by each enhancement method. To assess the significance of the observed differences in mean MUSHRA scores, a two-sample *t*-test was used with Satterthwaite's approximation for unequal variances [69].

The proposed method always has lower perceived reverberation than the unprocessed speech. It consistently achieves higher reverberation reduction than [14] and the difference in mean performance was statistically significant ($P < 5\%$). In most cases, the proposed method also improves on the quality of the unprocessed speech although, in a minority of cases, it appears that the strong reverberation and noise reduction applied by the algorithm leads to a small degradation in perceived quality. In most cases, the proposed method gave higher quality than [14] although the difference in mean improvement was not statistically significant at the 5% level.

From these results, the proposed algorithm is especially suited to situations with high levels of reverberation and/or noise. We believe that the algorithm is able to achieve large reductions in both noise and reverberation because it estimates them jointly rather than independently and also because its use of a speech model allows it to take advantage of correlations between frequency bands. In applications with lower levels of reverberation and noise, the method of [14] may be preferred since it has lower computational requirements and almost never degrades the perceived quality.

VI. Conclusion

In this paper, we have presented a novel blind single-channel approach to the online dereverberation problem. Using an ARMA model for the reverberation power and a Hidden Markov Model for the clean speech log-power, a spectral gain is computed in order to achieve good dereverberation performance. This real-valued gain is computed for each frame after jointly estimating posterior distributions of the acoustic parameters and speech, reverberation, and noise log-powers. The algorithm achieves very good dereverberation and denoising performance while improving speech quality and preserving speech intelligibility. Listening tests showed excellent audible quality of the speech signals processed by the proposed method.

Appendix A
Observation Noise

The complex STFT coefficients of the degraded speech observation can be modeled as zero-mean complex Gaussians in each time-frequency bin using the central limit theorem. Using $Y^{\circ}(l,\tilde{k})$ to denote the complex STFT coefficient of the observed speech at time frame l and at STFT frequency bin $\tilde{k}$, $Y^{\circ}(l,\tilde{k}) \sim \mathcal{N}\left(0, \sigma(l,\tilde{k})^2\right)$. We therefore have

$$|Y^{\circ}(l,\tilde{k})|^2 = \Re\{Y^{\circ}(l,\tilde{k})\}^2 + \Im\{Y^{\circ}(l,\tilde{k})\}^2 \tag{61}$$

where $\Re\{Y^{\circ}(l,\tilde{k})\}^2$ and $\Im\{Y^{\circ}(l,\tilde{k})\}^2$ are independent zero-mean Gaussians with variance $\frac{\sigma(l,\tilde{k})^2}{2}$. It follows that $\frac{|Y^{\circ}(l,\tilde{k})|^2}{\frac{\sigma(l,\tilde{k})^2}{2}} \sim \chi^2(2)$ or, equivalently,

$$|Y^{\circ}(l,\tilde{k})|^2 \sim \Gamma(1, \sigma(l,\tilde{k})^2). \tag{62}$$

As we formulated the problem in Mel-frequency bands, the power in STFT frequency bins of each time frame are then weighted and summed according to our filterbank. We assume the resulting weighted sum of Gamma distributed random variables is also approximately Gamma distributed, so that

$$\breve{\boldsymbol{Y}}_l(k) \sim \Gamma\left(\frac{1}{\kappa_k}, \kappa_k \sigma(l,k)^2\right) \tag{63}$$

with mean $\mathrm{E}[\breve{\boldsymbol{Y}}_l(k)] = \sigma(l,k)^2$ and variance $\mathrm{Var}[\breve{\boldsymbol{Y}}_l(k)] = \kappa_k\, \sigma(l,k)^4$. The values κ_k were determined empirically.

As we are assuming normally distributed log-powers, we can use the formula relating the moments of a normal distribution in the log-domain to the moments of a log-normal distribution in the power domain [70], and approximate the variance of $\boldsymbol{y}_l(k)$ as follows:

$$\mathrm{Var}[\boldsymbol{Y}_l(k)] \approx \log\left(1 + \frac{\mathrm{Var}[\breve{\boldsymbol{Y}}_l(k)]}{(\mathrm{E}[\breve{\boldsymbol{Y}}_l(k)])^2}\right) \tag{64}$$

$$= \log(1 + \kappa_k) \tag{65}$$

This means we have for the observation noise $\boldsymbol{\nu}_l \sim \mathcal{N}(0, \boldsymbol{M}_l)$ with $\boldsymbol{M}_l = \mathrm{diag}\left(\log(1+\boldsymbol{\kappa})\right)$ in (42).

Appendix B
Model Noise

As well as the observation noise that is a consequence of the statistical properties of the input data, we can model the noise due to the inaccuracies introduced when we assumed the powers are exactly additive. The total power in Mel-frequency band k

is therefore assumed to be

$$\begin{aligned}\breve{\boldsymbol{Y}}_l(k) &= \left|\sqrt{\breve{G}_l\breve{\boldsymbol{S}}_l(k)} + \sqrt{\breve{\boldsymbol{R}}_l(k)}e^{j\phi_k} + \sqrt{\breve{\boldsymbol{N}}_l(k)}e^{j\theta_k}\right|^2 \\ &= \breve{G}_l\breve{\boldsymbol{S}}_l(k) + \breve{\boldsymbol{R}}_l(k) + \breve{\boldsymbol{N}}_l(k) \\ &\quad + 2\sqrt{\breve{G}_l\breve{\boldsymbol{S}}_l(k)}\sqrt{\breve{\boldsymbol{R}}_l(k)}\cos(\phi_k) \\ &\quad + 2\sqrt{\breve{G}_l\breve{\boldsymbol{S}}_l(k)}\sqrt{\breve{\boldsymbol{N}}_l(k)}\cos(\theta_k) \\ &\quad + 2\sqrt{\breve{\boldsymbol{R}}_l(k)}\sqrt{\breve{\boldsymbol{N}}_l(k)}\cos(\theta_k - \phi_k)\end{aligned} \tag{66}$$

where θ_k and ϕ_k are respectively the phase differences between clean speech and noise, and clean speech and reverberation. If θ_k and ϕ_k are uniformly distributed in $[0, 2\pi]$ then $\theta_k - \phi_k$ is also uniformly distributed in $[0, 2\pi]$. It follows that the expectation of their cosine is 0, and the expectation of their squared cosine is $1/2$. We can therefore compute the moments of $\breve{\boldsymbol{Y}}_l(k)$, which gives:

$$\mathrm{E}[\breve{\boldsymbol{Y}}_l(k)] = \breve{G}_l\breve{\boldsymbol{S}}_l(k) + \breve{\boldsymbol{R}}_l(k) + \breve{\boldsymbol{N}}_l(k) \tag{67}$$

$$\begin{aligned}\mathrm{Var}(\breve{\boldsymbol{Y}}_l(k)) &= \mathrm{E}[\breve{\boldsymbol{Y}}_l(k)^2] - \mathrm{E}[\breve{\boldsymbol{Y}}_l(k)]^2 \\ &= 2\breve{G}_l\breve{\boldsymbol{S}}_l(k)\breve{\boldsymbol{R}}_l(k) + 2\breve{G}_l\breve{\boldsymbol{S}}_l(k)\breve{\boldsymbol{N}}_l(k) \\ &\quad + 2\breve{\boldsymbol{R}}_l(k)\breve{\boldsymbol{N}}_l(k)\end{aligned} \tag{68}$$

Using (64) we obtain the variance of the total log-power

$$\mathrm{Var}[\boldsymbol{Y}_l(k)] \approx \log\left(1 + 2\frac{\breve{G}_l\breve{\boldsymbol{S}}_l(k)\breve{\boldsymbol{R}}_l(k) + \breve{G}_l\breve{\boldsymbol{S}}_l(k)\breve{\boldsymbol{N}}_l(k) + \breve{\boldsymbol{R}}_l(k)\breve{\boldsymbol{N}}_l(k)}{\breve{G}_l\breve{\boldsymbol{S}}_l(k) + \breve{\boldsymbol{R}}_l(k) + \breve{\boldsymbol{N}}_l(k)}\right) \tag{69}$$

The observation noise covariance matrix $\boldsymbol{M}_l$ in (42) is therefore augmented by a diagonal matrix $\boldsymbol{T}_l$ whose diagonal elements are defined by (69), so that $\boldsymbol{M}_l = \mathrm{diag}\left(\log(1 + \boldsymbol{\kappa})\right) + \boldsymbol{T}_l$. This extra noise term is small when one of the powers is much greater than the others and maximum when all signal powers are equal (i.e. the point of maximum curvature of h).

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) DREAMS project under grant agreement ITN-GA-2012-316969 and from the Engineering and Physical Sciences Research Council under grant number EP/M026698/1. The authors would like to thank the anonymous reviewers for their helpful suggestions and Benjamin Cauchi for providing a MATLAB implementation of [14].

REFERENCES

[1] T. Houtgast, "The effect of ambient noise on speech intelligibility in classrooms," *Appl. Acoust.*, vol. 14, no. 1, pp. 15–25, 1981.

[2] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 331–342, 2006.

[3] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Hoboken, NJ, USA: Wiley, 2009.

[4] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation (ser. Signals and Communication Technology)*. New York, NY, USA: Springer-Verlag, 2010.

[5] K. Kinoshita *et al.*, "The REVERB challenge," Website, 2014. [Online]. Available: http://reverb2014.dereverberation.com/

[6] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Univ. Eindhoven, Eindhoven, the Netherlands, 2007.

[7] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[8] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 45–48.

[9] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multichannel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[10] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[11] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1617–1632, Jul. 2012.

[12] X. Xiao *et al.*, "The NTU-ADSC systems for reverberation challenge 2014," in *Proc. REVERB Challenge Workshop*, 2014, pp. o2.2:1–8.

[13] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoust.*, vol. 87, pp. 359–366, 2001.

[14] B. Cauchi *et al.*, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Adv. Signal Process.*, vol. 61, 2015, pp. 1–12.

[15] C. Evers and J. R. Hopgood, "Parametric modelling for single-channel blind dereverberation of speech from a moving speaker," *IET Signal Process.*, vol. 2, no. 2, pp. 59–74, Jun. 2008.

[16] A. Maezawa, K. Itoyama, K. Yoshii, and H. G. Okuno, "Nonparametric Bayesian dereverberation of power spectrograms based on infinite-order autoregressive processes," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1918–1930, Dec. 2014.

[17] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 276–289, Feb. 2016.

[18] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5177–5181.

[19] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 7, May 1982, pp. 1858–1861.

[20] B. D. Radlovic and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[21] K. Kinoshita *et al.*"A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 7, pp. 1–19, 2016.

[22] J.-D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, , Univ. du Maine, Le Mans, France, 1988.

[23] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[24] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 394–406, Feb. 2015.

[25] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 85–88.

[26] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, Mar. 1973.

[27] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[28] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[29] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[30] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 882–892, Mar. 2007.

[31] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[32] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB." [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2016.

[33] ITU-T, *Objective Measurement of Active Speech Level, International Telecommunication Union Recommendation P.56*, 1993.

[34] H. Kuttruff, *Room Acoustics*, 5th ed. London, U.K.: Spon Press, 2009.

[35] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 45–57, Sep. 1996.

[36] M. Gales and S. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Commun.*, vol. 12, pp. 231–239, 1993.

[37] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[38] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[39] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[40] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[41] S. Theodoridis, *Machine Learning—A Bayesian and Optimization Perspective*. New York, NY, USA: Academic, 2015.

[42] C. L. Thornton, "Triangular covariance factorizations for Kalman filtering," Ph.D. dissertation, California Inst. Technol., Pasadena, CA, USA, 1976.

[43] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*. New York, NY, USA: Academic, 1977.

[44] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 161–165.

[45] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab., Vol. 1: Statist.*, 1967, pp. 281–297.

[46] A. David and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Sympo. Discrete Algorithms*, 2007, pp. 1027–1035.

[47] B. Kulis and M. I. Jordan, "Revisiting k-means: New algorithms via Bayesian nonparametrics," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 513–520.

[48] N. Mohammadiha and A. Leijon, "Model order selection for non negative matrix factorization with application to speech enhancement," , KTH Roy. Inst. Technol., Stockholm, Sweden, *Tech. Rep.*, diva2:447310, 2011.

[49] W. Fisher, G. R. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specification and status," in *Proc. DARPA Speech Recognit. Workshop*, 1986, pp. 93–99.

[50] K. P. Burnham and D. R. Anderson, "Multimodel inference—Understanding AIC and BIC in model selection," *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, 2004.

[51] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2006, pp. A33:1–4.

[52] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Int. Conf. Digit. Signal Process.*, Jul. 2009, pp. 1–4.

[53] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 242–248, Feb. 1988.

[54] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Process.*, vol. 7, 1998, pp. 2819–2822.

[55] J. Y. C. Wen and P. A. Naylor, "An evaluation measure for reverberant speech using decay tail modelling," in *Proc. Eur. Signal Process. Conf.*, 2006, pp. 1–5.

[56] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 55–59.

[57] J. F. Santos, "MuSAELab—SRMR MATLAB toolbox," *Software Package*, 2014–2016. [Online]. Available: https://github.com/MuSAELab/SRMRToolbox

[58] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[59] C. Taal, "MATLAB code for algorithms," *Software Package*, 2011–2015. [Online]. Available: http://www.ceestaal.nl/matlab-code/

[60] ITU-T, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *International Telecommunication Union, Recommendation P.862*, 2000.

[61] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[62] S. Goetze *et al*. "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Sep. 2014, pp. 233–237.

[63] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[64] B. Cauchi, H. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. Naylor, "Perceptual and instrumental evaluation of the perceived level of reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 629–633.

[65] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—Corpus description and performance evaluation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[66] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *Proc. 110th AES Conv.*, Amsterdam, The Netherland, May 2001, pp. 12–15.

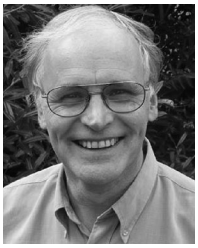
[67] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2005, pp. 357–362.

[68] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," *Radiocommunication sector of the International Telecommunication Union, Recommendation BS.1534–3*, 2001.

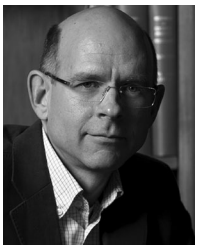
[69] G. W. Snedecor and W. G. Cochran, "The comparison of two samples," in *Statistical Methods*. Oxford, U.K.: Blackwell, 1989, ch. 6, pp. 83–106.

[70] N. L. Johnson, S. Kotz, and N. Balakrishnan, "Lognormal distributions," in *Continuous Univariate Distributions*. Hoboken, NJ, USA: Wiley, 1994, ch. 14.

Clement S. J. Doire (S'13) received the M.Sc. degree in communications and signal processing from Imperial College London, London, U.K., in 2012, the M.Eng. degree in electrical engineering from Ecole Supérieure d'Electricité, Gif-sur-Yvette, France, in 2013, and the Ph.D. degree from the Speech and Audio Processing Group, Imperial College London in 2016. During the Ph.D. degree, he was a fellow of the European Union Marie Curie Initial Training Network DREAMS project (Dereverberation and REverberation of Audio Music and Speech). His current research interests include speech enhancement, psychoacoustics, and machine learning. He received the Science and Communication Award for his involvement in the Royal Society's Summer Science Exhibition 2015.



Mike Brookes (M'88) received the Graduate degree in mathematics from Cambridge University, Cambridge, U.K., in 1972. He is currently a Reader (Associate Professor) in signal processing at the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. He was with the Massachusetts Institute of Technology and, briefly, with the University of Hawaii before returning to the U.K. and joining Imperial College in 1977. Within the area of speech processing, his work concentrates on the modeling and analysis of speech signals, the extraction of features for speech and speaker recognition, and on the enhancement of poor quality speech signals. He is the primary author of the VOICEBOX speech processing toolbox for MATLAB. Between 2007 and 2012, he was the Director of the Home Office sponsored Centre for Law Enforcement Audio Research, which investigated techniques for processing heavily corrupted speech signals. He is currently a Principal Investigator of the E-LOBES Project that seeks to develop environment-aware enhancement algorithms for binaural hearing aids.



Patrick A. Naylor (M'89–SM'07) received the B.Eng degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., in 1986 and the Ph.D. degree from Imperial College, London, U.K., in 1990. Since 1990, he has been a Member of Academic Staff, Department of Electrical and Electronic Engineering, Imperial College London. His research interests include the areas of speech, audio, and acoustic signal processing. He has worked, in particular, on adaptive signal processing for dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single and multichannel speech enhancement, and speech production modeling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the U.K., USA, and in mainland Europe. He is the Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the Director of the European Association for Signal Processing, and formerly an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and the IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING.

Christopher M. Hicks received the B.A., M.A., and Ph.D. degrees from the Signal Processing and Communications Group, Cambridge University Engineering Department, Cambridge, U.K. After the Ph.D. degree, he has been with CEDAR Audio, Ltd., where he is the Technical Director responsible for embedded DSP systems. In 2000, he was the Chair of the British Section of the Audio Engineering Society. In 2005, he received (with Dave Betts) the Technical Achievement Award by the Academy of Motion Picture Arts and Sciences for his work on the CEDAR DNS1000 film dialogue noise suppression system. He has been an Elected Fellow of Churchill College, Cambridge, since 2003.

Dave Betts received the B.A. degree in general engineering from the University of Cambridge, Cambridge, U.K. He was a founding employee of CEDAR Audio, Ltd., and is the Technical Director responsible for algorithm research and software development. In 2005, he received (with Dr. C. Hicks) the Technical Achievement Award by the Academy of Motion Picture Arts and Sciences for his work on the CEDAR DNS1000 film dialog noise suppression system.

Mohammad A. Dmour (S'04–M'10) received the B.Sc. degree in electrical engineering from the University of Jordan, Amman, Jordan, in 2005, and the M.Sc. (with distinction) degree in signal processing and communications and the Ph.D. degree in audio signal processing from the University of Edinburgh, Edinburgh, U.K., in 2006 and 2010, respectively. Since 2010, he has been a Research Engineer at CEDAR Audio, Ltd., Cambridge, U.K. He received the Wolfson Microelectronics Scholarship for the pursuit of the Ph.D. degree at the University of Edinburgh, and was presented with the class medal upon the culmination of the M.Sc. degree in signal processing and communications. His current research interests include audio source separation and speech enhancement.

Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. He is currently a Full Professor at Aalborg University. Before joining the Department of Electronic Systems, Aalborg University, he was with the Telecommunications Laboratory, Telecom Denmark, Ltd., Copenhagen, Denmark; the Electronics Institute of Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education, Lyngby; the Electrical Engineering Department, Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation, Aalborg University. His current research interests include numerical algorithms, optimization engineering, machine learning, and digital processing of acoustic, audio, communication, image, multimedia, speech, and video signals. He is the coauthor of the textbook *Software-Defined GPS and Galileo Receiver—A Single-Frequency Approach* (Springer, 2007), also translated to Chinese: National Defence Industry Press, China. He has been the Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, *Elsevier Signal Processing*, and the *EURASIP Journal on Advances in Signal Processing*. He received an individual European Community Marie Curie Fellowship, the former Chairman of the IEEE Denmark Section, and the IEEE Denmark Sections Signal Processing Chapter. He is the Member of the Danish Academy of Technical Sciences and has been a member of the Danish Council for Independent Research appointed by the Danish Minister of Science.