ARTICULATORY BASED SPEECH MODELS FOR BLIND SPEECH DEREVERBERATION USING SEQUENTIAL MONTE CARLO METHODS

Christine Evers and James R. Hopgood

School of Engineering and Electronics, Institute for Digital Communications, The University of Edinburgh
Mayfield Road, Edinburgh, EH9 3JL, United Kingdom
phone: +44 (131 650) 5655, email: c.evers@ieee.org
web: www.see.ed.ac.uk, http://www.erp.ac.uk/

ABSTRACT

Room reverberation leads to reduced intelligibility of audio signals. Enhancement is thus crucial for high-quality audio and scene analysis applications. This paper proposes to directly and optimally estimate the source signal and acoustic channel from the distorted observations. The remaining model parameters are sampled from a particle filter, facilitating real-time dereverberation. The approach was previously successfully applied to single- and multi-sensor blind dereverberation. Enhancement can be improved upon by accurately modelling the speech production system. This paper therefore extends the blind dereverberation approach to incorporate a novel source model based on parallel formant synthesis and compares the approach to one using a time-varying AR model, with parameters varying according to a random walk. Experimental data shows that dereverberation using the proposed model is improved for vowels, stop consonants, and fricatives.

1. INTRODUCTION

Room reverberation leads to reduced intelligibility of audio signals and spectral coloration of audio signals. Thus, for high quality of digitally recorded speech, blind dereverberation of the observed signal is crucial in order to obtain an anechoic speech estimate [1–3].

The problem of source signal estimation could be considered from a maximum-likelihood (ML) perspective. However, if the production mechanism and distorting environment are unknown, the ML approach of source signal estimation requires the maximisation of the likelihood over any parameters specifying the source production mechanism, the distorting channel, and, above all, the source signal with respect to the known observations. Therefore, without any available prior knowledge of the underlying production mechanism, an infinite parameter space is to be searched.

It can therefore prove highly advantageous to incorporate prior information about the source production mechanism and distorting channel in the estimation process. As exact knowledge of the vocal tract and room transfer function are generally unavailable, models of the vocal tract and room acoustics are utilised instead. Estimates can therefore be improved upon by accurate source modelling.

This paper proposes a novel speech model based on a parallel formant synthesiser (PFS). PFSs model the formants of speech by a parallel concatenation of several resonant circuits. Each circuit is represented by a second-order autoregressive (AR) process and is driven by an amplitude control, setting the resonant frequency and bandwidth, i.e., the height and width of the formants' spectral peaks. In practice, the resonant frequencies and bandwidths are unknown and therefore need to be modelled as well. In order to account for the time-varying properties of speech, the frequency and bandwidth of each resonator could be allowed to vary according to a random walk. However, unbounded sampling does not necessarily enforce frequencies between 0 and π. Therefore, this paper investigates alternative parameterisation of the AR process in order to facilitate valid frequencies and bandwidths, whilst ensuring stable AR parameters. It is proposed to parameterise the PFS in terms of partial correlation (PARCOR) coefficients whose values correspond to valid bounded resonant frequencies and bandwidths.

The proposed model is compared to a time-varying AR (TVAR) source model, where the TVAR parameters are assumed to vary according to a random walk. Both models are implemented within a blind dereverberation approach efficiently applied previously in, e.g., [4, 5]. In this framework, the source signal and reverberant channel are obtained using their optimal estimator, the Kalman filter, whereas the remaining model parameters are estimated by sequential importance resampling.

This paper is structured as follows: Sect. §2 introduces the general system model, sect. §3 discusses the TVAR source model and derives the proposed PFS model, sect. §4 derives the blind speech dereverberation algorithm, and sect. §5 compares the performance of the blind dereverberation approach for both source models based on speech data.

2. GENERAL SYSTEM MODEL

The speech production mechanism can be modelled as a concatenation of lossless acoustic tubes of equal lengths, whose transfer function can be approximated by an all-pole filter [6]. Furthermore, the solution of the acoustic wave function suggests that the transfer function of geometric reverberant rooms can be modelled as an all-pole filter. The source and observed signal can therefore be easily expressed in state-space form as

$$\mathbf{x}_t = \mathrm{A}_t \mathbf{x}_{t-1} + \Sigma_{\mathbf{v}_t} \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}_{Q\times 1}, \mathrm{I}_Q\right), \quad (1a)$$

$$\mathbf{y}_t = \mathrm{Y}_{t-1}\mathbf{b} + \mathrm{C}^T \mathbf{x}_t + \Sigma_{w_t} \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}_{M\times 1}, \mathrm{I}_M\right), \quad (1b)$$

where $\mathbf{x}_t = [x_t \quad \dots \quad x_{t-Q+1}]^T$ are the most recent Q source signal samples, $\mathbf{a}_t = [a_{1,t} \quad \dots \quad a_{Q,t}]^T$ are the source parameters, A_t is the source transition matrix governed by the source model parameters and $\Sigma_{\mathbf{v}_t}$ is the covariance matrix of the source excitation, $\mathbf{v}_t$. $\mathbf{y}_t = [y_{1,t} \quad \dots \quad y_{M,t}]^T$ are the M sensor observations, $\mathbf{b} = \left[\mathbf{b}_1^T \quad \dots \quad \mathbf{b}_M^T\right]^T$ where $\mathbf{b}_m = [b_{1,m} \quad \dots \quad b_{P,m}]^T$ are the P channel parameters between the source and the m^{th} sensor, and $\mathrm{Y}_{t-1} = \mathrm{diag}\left[\widehat{\mathbf{y}}_{1,t-1}^T \quad \dots \quad \widehat{\mathbf{y}}_{M,t-1}^T\right]$ contains the P past samples at each sensor, where $\widehat{\mathbf{y}}_{m,t-1} \triangleq [y_{m,t-1} \quad \dots \quad y_{m,t-P}]^T$. Furthermore, $\mathrm{C}^T = \mathbf{1}_{M\times 1}\mathbf{c}^T$, where $\mathbf{c}^T$ is a $1\times Q$ source-model dependent combination of ones and zeros retaining only the samples of $\mathbf{x}_t$ required for the generation of $\mathbf{y}_t$. Note that a distinct white Gaussian noise (WGN) noise source, $\mathbf{w}_t$, with $M\times M$ covariance matrix Σ_{w_t}, and close to the target source is incorporated by a simplification of the common-acoustical pole and zero (CAPZ) model [7] as discussed in, e.g., [8].

3. SPEECH MODELS

3.1 Markov chain based TVAR model

The TVAR parameters vary slowly and relatively smoothly. The smooth and slowly varying behaviour can be represented by a first-order Markov chain with low variance on the parameters, i.e.,

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \Sigma_{\mathbf{a}}\, \mathbf{r}_{\mathbf{a}_t} \qquad \mathbf{r}_{\mathbf{a}_t} \sim \mathcal{N}\left(\mathbf{0}_{Q\times 1}, \mathrm{I}_Q\right) \quad (2)$$

 ISSN 2076-1465

where $\Sigma_{\mathbf{a}} = \text{diag}\left[\sigma^2_{a_{1,t}} \quad \dots \quad \sigma^2_{a_{Q,t}}\right]$ is assumed known.

An issue encountered with the Markov chain model is the necessity to constrain the parameters in eqn. (2) to take on stable values only, i.e., to have poles within the unit circle. Stability can be enforced by reflecting unstable poles back into the unit circle [9] by letting $p_{q,t} = 1 \Big/ \hat{p}_{q,t}$, where $\{p_{q,t}\}_{q\in\mathscr{Q}}$ are the Q poles corresponding to the roots of $\mathbf{a}_t$ and $\hat{p}_{q,t}$ denotes a pole outside the unit circle. As poles appear in complex-conjugate pairs, the reflection changes the radius but leaves the phase unchanged.

The Markov chain based speech model can be improved upon by exploiting physical descriptions of the human vocal tract.

3.2 Novel parallel formant synthesis TVAR model

PFSs model the formants (or spectral peaks) generated in the vocal tract as a parallel concatenation of K resonators, each of which models one formant and is preceded by an amplitude control of the spectral peak. Each resonator signal can be modelled by a second-order TVAR process with a complex-conjugate pair located inside and close to the unit circle, where, for each resonator $k \in \mathscr{K}$,

$$x_{k,t} = a_{1,t,k}\, x_{k,t-1} + a_{2,t,k}\, x_{k,t-2} + \sqrt{g_{k,t}}\, v_t, \quad v_t \sim \mathscr{N}(0,1). \quad (3)$$

where $\{x_{k,t} : t \geq Q, k \in \mathscr{K}\}$ is the resonator output, $\{a_{q,t,k} : q \in \mathscr{Q}, k \in \mathscr{K}\}$ are the source parameters, and $g_{k,t}$ is the resonator gain. The K resonator signals are combined to form the synthetic speech signal as $x_t = \sum_{k\in\mathscr{K}} x_{k,t}$.

The main concern for designing the resonators is to ensure poles located near the unit circle to generate large magnitude responses at the desired positions in the spectrum. The TVAR parameters are therefore specified by design criteria characterising constraints on the frequency response, $H_{k,t}(\omega)$,

$$H_{k,t}(\omega) = \frac{g_{k,t}}{\left(1 - r_{k,t}\, e^{j\phi_{k,t}}\, e^{-j\,\omega}\right)\left(1 - r_{k,t}\, e^{-j\phi_{k,t}}\, e^{-j\,\omega}\right)} \quad (4)$$

where $p_{1,t,k} = r_{k,t} e^{j\phi_{k,t}}$ and $p_{2,t,k} = p^{\star}_{1,t,k} = r_{k,t} e^{-j\phi_{k,t}}$ are the two poles of the filter, where $\cdot^{\star}$ denotes the complex conjugate, $r_{k,t}$ is the pole radius, $\phi_{k,t}$ is the pole phase, $\omega = {}^{2\pi f}/_{f_s}$ denotes the radial frequency, and f_s is the sampling frequency. The pole radius and phase can be related to the TVAR parameters for $Q = 2$ via [10]

$$a_{1,t,k} = -2 r_{k,t} \cos\phi_{k,t} \qquad a_{2,t,k} = r^2_{k,t}. \quad (5)$$

The most crucial design criteria for PFSs are the specification of the resonant frequency and 3dB bandwidth of the resonator, i.e.,

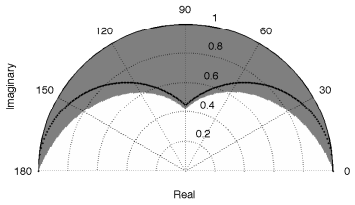
$$\frac{\partial}{\partial\omega}|H_{k,t}(\omega)|\Big|_{\omega=\omega_{k,t}} = 0 \quad \text{and} \quad |H_{k,t}(\omega)|\Big|_{\omega=\omega_{k,t}\pm B_{k,t}/2} = G_B$$

where G_B is the gain at the 3dB bandwidth, $B_{k,t}$ is the 3dB bandwidth, and $\omega_{k,t}$ is the radial frequency at resonance. Inserting eqn. (4) and solving for $\omega_{k,t}$ and $B_{k,t}$ respectively, $f_{k,t}$ and $B_{k,t}$ can be related to $r_{k,t}$ and $\phi_{k,t}$ [9]. As the resonant frequency and bandwidth are related to the poles, and the poles are related to the TVAR parameters, $\mathbf{a}_{k,t}$ can be related to $f_{k,t}$ and $B_{k,t}$.

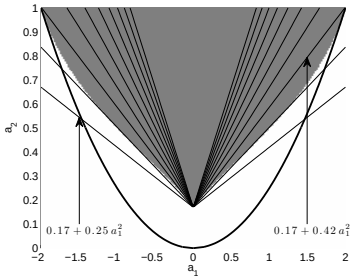
In order to obtain estimates of the TVAR parameters in an sequential importance resampling (SIR) framework, it seems tempting to model the formant frequency, bandwidth and gain as a random walk similar to eqn. (2) [11]. However, the bandwidth and resonant frequency must be limited by $0 \leq f_{k,t} \leq \pi$ and $0 \leq B_{k,t} \leq \pi$ whilst the TVAR parameters, $\mathbf{a}_{k,t}$ must have poles within the unit circle. These constraints cannot be enforced by unbounded sampling from $f_{k,t}$ and $B_{k,t}$. It is therefore of interest to investigate the region of parameters corresponding to both valid AR parameters and resonant frequencies / bandwidths.

3.3 Admissible regions of parameters

In order to identify the region of stable and valid AR parameters, a grid of 200×200 poles is generated within the unit circle, i.e., with pole radius $0 \leq r_t \leq 1$ and phase $0 \leq \phi_t \leq \pi$. For each pole, the magnitude response, $|H_t(j\omega)|$, is evaluated and it is tested whether the spectral peak is sufficiently high for the extraction of the 3dB bandwidth. The region of stable poles corresponding to valid resonant frequencies and 3dB bandwidths is identical for both experiments and displayed as a grey shape in the pole space in Fig. 1a and in the parameter space in Fig. 1b.



(a) Poles corresponding to Fig. 1b.



(b) Triangle approximation in parameter space.

Figure 1: Areas of stable parameters and valid resonant frequencies / bandwidths in parameter and pole space (grey areas) vs. approximations in parameter space (black lines).

3.3.1 *Approximation in parameter space*

Due to the unusual shape of the valid and stable regions in both the pole and parameter space, an exact description of the regional shape is not obvious and approximations are necessary. The valid region of parameters in Fig. 1a resembles a hybrid shape between a triangle and an ellipsoid. One would therefore expect to approximate the boundaries of this shape either by an ellipse or a triangle. However, ellipses were found to be an unsuitable approximation.

Instead of an ellipse, the shape in Fig. 1b is therefore best approximated using an isosceles triangle where $\max\{a_{2,t}\} = 1 - 0.17 = 0.83$. The gradient of the triangle is therefore ${}^{\max\{a_{2,t}\}}/_{\max\{a_{1,t}\}} \approx 0.41$:

$$a_{2,t} = 0.17 \pm 0.41\, a_{1,t}. \quad (6)$$

Fig. 1b verifies these results by comparing the fit of the triangle in eqn. (6) to triangles with increasing gradients between $\alpha = {}^1/_4, \dots, 1$. The triangle specified in eqn. (6) omits the smallest portion of the valid regions and avoids the inclusion of invalid areas. However, the transforming the shape to pole space, indicated as a black line in Fig. 1a, a relatively large proportion of valid resonant poles close to the unit circle are excluded.

3.3.2 *Approximation in pole space*

In order to reduce the number of resonant and valid poles excluded from the approximated region of support, the valid and stable region can be approximated directly in the z-domain rather than the parameter space. Again, due to the shape of the region of support in Fig. 1a, an exact description of the boundaries seems non-obvious. Therefore, an ellipse is used for approximation, i.e., $\phi_t = \max\{\phi_t\}\sqrt{1 - r_t^2}$. where the imaginary part is normalised between $0 \leq \phi_t \leq 1$. The most accurate approximation is achieved for $\max\{\phi_t\} = 0.5875$ (see Fig. 2). Although the ellipse fails to model the lobe between $120 \leq \phi_t \leq 60$ and $0.4 \leq r_t \leq 0.6$, the magnitude responses are comparatively flat in this region due to its distance from the unit circle and proximity to the origin. Therefore,

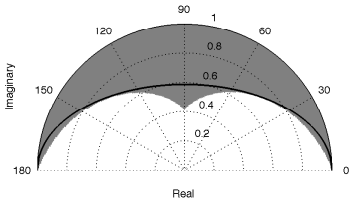


Figure 2: Grey areas correspond to regions of stable parameters generating valid 3dB bandwidths for max $\{0.5875\}$.

the poles in the lobe only add a minor contribution to the frequency response. However, the approximated region of support in Fig. 2 still excludes a small portion of resonant poles between $0 \leq \phi_t \leq 40$ and $180 \leq \phi_t \leq 140$. An approximation of the region of support is thus desirable excluding the central lobe of the hour-glass shape, whilst including any valid areas near the unit circle.

Instead of parameterising eqn. (3) in terms of AR coefficients or poles using a direct-form infinite impulse response (IIR) structure, the model can be represented by a lattice IIR structure and be parameterised in terms of the lattice reflection coefficients [9]. The reflection coefficients of an IIR lattice structure correspond to so-called PARCOR coefficients, describing the relation between the forward and backward lattice structure [9]. This description is directly related to the relation between the propagated and reflected sound waves at junctions in the acoustic tube, such that the reflection, or PARCOR, coefficients of the lattice structure are equivalent to the reflection coefficients of the vocal tract transfer function. The PARCOR interpretation of TVAR models is thus a popular alternative to the AR parameters [6, 9] and offers an interesting alternative for investigating the region of valid parameters and resonant frequencies / bandwidths.

3.4 Approximation in PARCOR space

The TVAR parameters are related to the PARCOR coefficients, $\{\psi_{q,t}\}_{q\in\mathscr{Q}}$, for a second-order model via [9]

$$a_{1,t} = \psi_{1,t}\left(1+\psi_{2,t}\right) \qquad \text{and} \qquad a_{2,t} = \psi_{2,t}. \tag{7}$$

Similar to the approximation in pole space, the area of stable AR parameters corresponding to valid resonant frequencies and 3dB bandwidths can therefore be reflected into the PARCOR coefficient domain using eqn. (7). The resulting region is shown as a grey shape in Fig. 3a. The shape resembles a full-bodied ellipse with a triangular peak. An ellipse using $\max\{\psi_{1,t}\} = 1$ is therefore fitted to the region of support in PARCOR parameter space, where

$$\psi_{2,t} = 1 - \max\{\psi_{2,t}\}\sqrt{1-\psi_{1,t}^2}. \tag{8}$$

Fig. 3a shows the ellipses for $\max\{\psi_{2,t}\} = 0,\ldots,2/3$ as black lines where $\max\{\psi_{2,t}\} = 2/3$ is the most accurate approximation.

The elliptical PARCOR approximation in pole space omits the central lobe between $60 \leq \phi_t \leq 120$ and $0.4 \leq r_t \leq 0.6$ similar to Fig. 2. Nonetheless, the PARCOR approximation does not exclude any resonant poles close to the unit circle. Therefore, the approximation in PARCOR parameter space provides the most accurate approximation of the valid regions as compared to the approximation in pole space or parameter space in Figures 3.3 and 2.

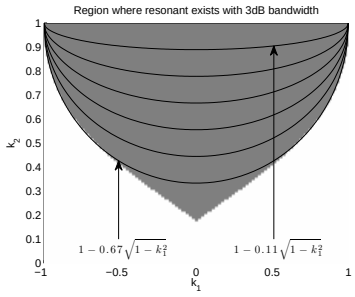
Rather than modelling the resonant frequency and 3dB bandwidth as a random walk, valid frequencies and bandwidths can be ensured by modelling the PARCOR coefficients as a random walk and reflecting the samples into the area in Fig. 3.

3.5 Reflection of PARCOR coefficients into valid region

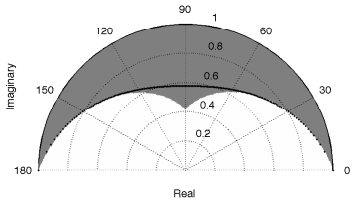
The PARCOR coefficients are therefore assumed to vary via

$$\widehat{\psi}_{t,k} = \psi_{t-1,k} + \Sigma_{\psi_{t,k}}\,\mathbf{r}_{\psi_{t,k}} \tag{9}$$

Instead of rejecting unstable or invalid samples, it is proposed to utilise bounded functions to transform *all* samples into the ellipsoid area in Fig. 3. Any bounded function, e.g., inverse trigonometric



(a) Parameter space



(b) Pole space

Figure 3: Grey areas correspond to regions of stable parameters generating valid 3dB bandwidths.

functions such as the arctan or arcsin, can be used. In this paper, the inverse logit function is employed, where

$$\upsilon = \text{logit}^{-1}(\chi) = \frac{1}{1+e^{-\chi}} \tag{10}$$

where $0 \leq \upsilon \leq 1$ for any $-\infty \leq \chi \leq \infty$. The stable and valid area of PARCOR coefficients in Fig. 3a is bounded between $-1 \leq \psi_{1,t} \leq 1$ and $1 \leq \psi_{2,t} \leq 1 - \frac{2}{3}\sqrt{1-\psi_{1,t}^2}$, whereas the inverse logit is defined between 0 and 1. Eqn. (10) is thus shifted scaled, such that

$$\psi_{1,t,k} = -1 + \frac{2}{1+\exp\{-\widehat{\psi}_{1,t,k}\}} \tag{11a}$$

$$\psi_{2,t,k} = \alpha + \frac{1-\alpha}{1+\exp\{-\widehat{\psi}_{2,t,k}\}} \tag{11b}$$

Therefore, as $\psi_{0:t}$ is enforced to lie in the stable and valid region of support shown in Fig. 3, valid resonant frequencies are ensured with resonant peaks of valid 3dB bandwidths. Furthermore, only one transformation from the PARCOR to the AR parameter space is necessary, rather than several transformations between the resonant frequencies and bandwidths, the AR parameter and pole space. The proposed sampling scheme therefore facilitates a simplified sampling scheme avoiding multiple transformations between parameter spaces and ensuring stability and valid frequencies / bandwidths.

4. METHODOLOGY

Given the stochastic model in eqn. (1), a sequential optimal estimator is sought of the unknown source signal at time t, $\mathbf{x}_t$. As the source signal is to be estimated blindly, it is necessary to estimate all variables, $\varphi_t \triangleq \begin{bmatrix}\mathbf{x}_t^T & \mathbf{b}^T & \theta_t^T\end{bmatrix}^T$ in order to obtain an estimate of $\mathbf{x}_t$, where $\theta_t \triangleq \{\mathbf{a}_t, \phi_{v_t}, \phi_{\mathbf{w}_t}\}$. If the unknown variables are considered as stochastic quantities, their estimate, $\hat{\varphi}_t$, can be obtained by their minimum mean-square error (MMSE) estimator, i.e.,

$$\begin{aligned}\hat{\varphi}_t &= \int \varphi_{0:t}\, p(\varphi_t \mid \mathbf{y}_{1:t})\, d\varphi_t \\ &= \iint \begin{bmatrix}\mathbf{z}_t \\ \theta_t\end{bmatrix} p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \theta_t)\, p(\theta_t \mid \mathbf{y}_{1:t})\, d\mathbf{z}_t\, d\theta_t \\ &= \begin{bmatrix}\int \hat{\mathbf{z}}_t\, p(\theta_t \mid \mathbf{y}_{1:t})\, d\theta_t \\ \hat{\theta}_t\end{bmatrix}\end{aligned} \tag{12}$$

where $\hat{\mathbf{z}}_t$ is the MMSE estimate of $\mathbf{z}_t = \begin{bmatrix}\mathbf{x}_t^T & \mathbf{b}^T\end{bmatrix}^T$, and $\hat{\theta}_{0:t}$ is the MMSE estimate of θ_t, and where, similar to eqn. (12),

$$\hat{\mathbf{z}}_t = \begin{bmatrix} \int \mathbf{x}_t \, p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}, \theta_{0:t}) \, d\mathbf{x}_t \\ \int \mathbf{b} \, p(\mathbf{b} \,|\, \mathbf{y}_{1:t}, \theta_{0:t}) \, d\mathbf{b} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}_t \\ \widehat{\mathbf{b}} \end{bmatrix} \tag{13}$$

where $p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}, \theta_t) = \int p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}, \theta_t, \mathbf{b}) \, p(\mathbf{b} \,|\, \mathbf{y}_{1:t}, \theta_t) \, d\mathbf{b}$ *marginalises* the channel parameters from the source signal posterior probability density function (pdf).

Therefore, estimates of $\mathbf{x}_t$, θ_t, and $\mathbf{b}$ can be obtained using three separate estimators. Having obtained $\hat{\theta}_t$, the results are used to estimate $\mathbf{b}$. Using $\widehat{\mathbf{b}}$ and $\hat{\theta}_{0:t}$, the source signal is estimated. Sect. §4.1 to sect. §4.4 therefore derive the three estimators.

4.1 Parameter estimation using particle filtering

For most speech parameter models, the posterior pdf of θ_t, required to solve eqn. (12) cannot derived in closed form. Therefore, θ_t cannot be estimated analytically. Instead, as an exercise in *stochastic* integration, Monte Carlo sampling can be used to approximate $\hat{\theta}_t$ by drawing N independent and identically distributed samples, $\theta_t^{(i)}, i \in \mathscr{N}$ from a hypothesis distribution that approximates and has the same support as the posterior pdf, $p(\theta_t \,|\, \mathbf{y}_{1:t})$. Each samples (or 'particle'), $\theta_t^{(i)}$ is associated with a weight proportional to its likelihood. The MMSE estimate can therefore be expressed as the point-mass distribution:

$$\hat{\theta}_t = \frac{1}{N} \sum_{i \in \mathscr{N}} \theta_t^{(i)} \tilde{w}_t^{(i)} \Big/ \sum_{j \in \mathscr{N}} \tilde{w}_t^{(j)},$$

where the importance weights are given as

$$w_t^{(i)} = w_{t-1}^{(i)} \, p\left(\mathbf{y}_{1:t} \,|\, \theta_t^{(i)}\right) p\left(\theta_t \,|\, \theta_{t-1}^{(i)}\right) \Big/ \pi\left(\theta_t^{(i)} \,|\, \mathbf{y}_{1:t}\right) \tag{14}$$

and are normalised via

$$\tilde{w}_t^{(i)} \triangleq w_t^{(i)} \Big/ \sum_{j \in \mathscr{N}} w_t^{(j)}. \tag{15}$$

The performance of particle filters is highly dependent on the choice of the hypothesis distribution, $\pi(\theta_t \,|\, \mathbf{y}_{1:t})$. The optimal importance function minimises the variance upon $\theta_t^{(i)}$ and the observations. However, generally $\theta_t^{(i)}$ are non-linear in the likelihood and $w_t^{(i)}$ cannot be evaluated. Sampling from the prior, $p(\theta_t \,|\, \theta_{t-1})$, is used in this paper, such that eqn. (14) reduces to

$$w_t^{(i)} = w_{t-1}^{(i)} \, p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_t^{(i)}\right). \tag{16}$$

Furthermore, as $\pi(\theta_t \,|\, \mathbf{y}_{1:t})$ only approximates $p(\theta_t \,|\, \mathbf{y}_{1:t})$ and the discrepancy increases with time, after few iterations all but one importance weight are close to zero and computational effort is dissipated to tracking particle trajectories not contributing to the final estimate. Resampling ensures that only statistically relevant samples are retained.

For each of the sampled parameters, the channel and source signal are to be estimated according to eqn. (12). Therefore, for each choice of $\theta_t^{(i)}$, an estimate of $\mathbf{b}^{(i)}$ and $\mathbf{x}_t^{(i)}$ is obtained using the estimators described in the following. Note that the superscript (i) is dropped for brevity.

4.2 Source signal estimation using the Kalman filter (KF)

The Kalman filter is the optimal estimator of the source signal for known model parameters, $\theta_{0:t}$, in conditionally Gaussian state-space (CGSS) systems such as eqn. (1). KFs sequentially predict $\mathbf{x}_{0:t}$ based on the model parameters and correct the prediction using the most recent measurement. The KF equations are found by 1) predicting the states based on previous data only, i.e., $p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t-1}, \theta_t, \mathbf{b})$ and 2) updating the estimate using y_t by applying Bayes's theorem, i.e., $p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}, \theta_t, \mathbf{b})$. Similar to [12] the Kalman equations for the reverberant state space are given as

$$\mu_{t|t-1} = \mathrm{A}_t \mu_{t-1|t-1}, \tag{17a}$$

$$\Sigma_{t|t-1} = \Sigma_{\mathbf{v}_t} \Sigma_{\mathbf{v}_t}^T + \mathrm{A}_t \Sigma_{t-1|t-1} \mathrm{A}_t^T \tag{17b}$$

$$\mu_{t|t} = \left(\mathrm{I}_Q - \mathrm{K}_t \mathrm{C}^T\right) \mu_{t|t-1} - \mathrm{K}_t \left(\mathrm{Y}_{t-1}\mathbf{b} - \mathbf{y}_t\right) \tag{17c}$$

$$\Sigma_{t|t} = \left(\mathrm{I}_Q - \mathrm{K}_t \mathrm{C}^T\right) \Sigma_{t|t-1}., \tag{17d}$$

with residual variance is $\Sigma_{z_t} = \Sigma_{w_t} \Sigma_{w_t}^T + \mathrm{C}^T \Sigma_{t|t-1} \mathrm{C}$, and Kalman gain is $\mathrm{K}_t = \Sigma_{t|t-1} \mathrm{C} \Sigma_{z_t}^{-1}$. The likelihood of the observations is

$$p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_t, \mathbf{b}) = \mathscr{N}\left(\mathbf{y}_t \,\middle|\, \mathrm{Y}_{t-1}\mathbf{b} + \mathrm{C}^T \mu_{t|t-1}, \Sigma_{z_t}\right). \tag{18}$$

The source signal can be estimated using its optimal estimator. However, both $p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}, \theta_t, \mathbf{b})$ and $p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_t, \mathbf{b})$ are still dependent on $\mathbf{b}$, which is unknown in practice.

4.3 Channel estimation using the KF

The static IIR component, $\mathbf{b}$, does not exhibit a dynamic over time. Predicting future values would thus be futile. Nonetheless, *belief* in the static parameters can be updated as new data becomes available. Using Bayes's theorem, this belief can be sequentially updated via

$$p(\mathbf{b} \,|\, \mathbf{y}_{1:t}, \theta_{0:t}) = \frac{p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}, \mathbf{b}) p(\mathbf{b} \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1})}{p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t})},$$

where the posterior pdf at time $t-1$, $p(\mathbf{b} \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1})$, acts as the prior pdf at t to recursively update $p(\mathbf{b} \,|\, \mathbf{y}_{1:t}, \theta_{0:t})$. Assuming that the posterior at $t-1$ is Gaussian with mean $\mu_{\mathbf{b},t-1}$ and covariance $\Sigma_{\mathbf{b},t-1}$,

$$p(\mathbf{b} \,|\, \mathbf{y}_{1:t}, \theta_{0:t}) = \mathscr{N}\left(\mathbf{b} \,\middle|\, \mu_{\mathbf{b},t}, \Sigma_{\mathbf{b},t}\right), \tag{19}$$

where the covariance, $\Sigma_{\mathbf{b},t}$, and mean, $\mu_{\mathbf{b},t}$, are given by

$$\mu_{\mathbf{b},t} = \left(\mathrm{I}_{MP} - \mathrm{K}_{\mathbf{b},t} \tilde{\mathrm{Y}}_{t-1}^T\right) \mu_{\mathbf{b},t-1} + \mathrm{K}_{\mathbf{b},t} \tilde{\mathbf{y}}_t \tag{20a}$$

$$\Sigma_{\mathbf{b},t} = \left(\mathrm{I}_{MP} - \mathrm{K}_{\mathbf{b},t} \tilde{\mathrm{Y}}_{t-1}^T\right) \Sigma_{\mathbf{b},t-1}, \tag{20b}$$

where $\tilde{\mathrm{Y}}_{t-1}^T \triangleq \mathrm{Y}_{t-1} + \mathrm{C}^T \Gamma_{t|t-1}$ and $\mathrm{K}_{\mathbf{b},t} = \Sigma_{\mathbf{b},t-1} \tilde{\mathrm{Y}}_{t-1} \Sigma_{\mathbf{z}_t,\mathbf{b}}^{-1}$ and $\Sigma_{\mathbf{z}_t,\mathbf{b}} = \Sigma_{z_t} + \tilde{\mathrm{Y}}_{t-1}^T \Sigma_{\mathbf{b},t-1} \tilde{\mathrm{Y}}_{t-1}$. Comparing eqn. (20) to eqn. (17), the channel estimation is of the form of the update Kalman equations. As more knowledge about the observations becomes available, the *belief* in the static IIR component is updated (as opposed to predicting a dynamic into the future and correcting using measurements as in sect. §4.2).

4.4 Marginalization of channel parameters

The Kalman equations for $\mathbf{x}_{0:t}$ are dependent on the channel parameters through $\mu_{t|t}$ (eqn. (17c)). In fact, as can be shown by induction, $\mu_{t|t}$ is *linearly dependent* in $\mathbf{b}$, such that eqn. (17c) at $t-1$ is equivalent to,

$$\mu_{t-1|t-1} = \underbrace{\mu_{t-1|t-2} + \mathrm{K}_{t-1}\left(\mathbf{y}_{t-1} - \mathrm{C}^T \mu_{t-1|t-2}\right)}_{\alpha_{t-1|t-1}} \underbrace{-\mathrm{K}_{t-1}\mathrm{Y}_{t-2}}_{\Gamma_{t-1|t-1}} \mathbf{b}$$

Inserting into the prediction in eqn. (17a) at t,

$$\mu_{t|t-1} = \mathrm{A}_t \mu_{t-1|t-1} = \underbrace{\mathrm{A}_t \alpha_{t-1|t-1}}_{\alpha_{t|t-1}} + \underbrace{\mathrm{A}_t \Gamma_{t-1|t-1}}_{\Gamma_{t|t-1}} \mathbf{b}. \tag{21}$$

Thus, $\mu_{t|t-1}$ is *implicitly linear* in $\mathbf{b}$ via $\mu_{t-1|t-1}$. Inserting eqn. (21) in eqn. (17c) and defining $\mathrm{B}_t \triangleq \mathrm{I}_Q - \mathrm{K}_t \mathrm{C}^T$, the update equation is thus linear in $\mathbf{b}$ through the relation

$$\mu_{t|t} = \underbrace{\mathrm{B}_t \alpha_{t|t-1} + \mathrm{K}_t \mathbf{y}_t}_{\alpha_{t|t}} + \underbrace{\left[\mathrm{B}_t \Gamma_{t|t-1} - \mathrm{K}_t \mathrm{Y}_{t-1}\right]}_{\Gamma_{t|t}} \mathbf{b}$$

This linear dependency of $\mu_{t|t}$ in $\mathbf{b}$ facilitates marginalization of $\mathbf{b}$ from $p(\mathbf{x}_t \,|\, \mathbf{y}_{1:t}, \theta_t, \mathbf{b})$ as shown in eqn. (13).

	Vowels	Stops	Fricatives	Semivowels
Observed	-3.46	-1.90	-2.99	-4.47
PFS	**+3.32**	**+7.86**	**+7.02**	-1.52
Markov chain	-1.23	+4.53	+2.97	**+0.45**

Table 1: Comparison of SNR in dB for PFS and Markov chain based models for different phoneme types.

4.4.1 *Marginalization of channel from state posterior*

Recalling $p(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \theta_t) = \int p(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \theta_t, \mathbf{b})\, p(\mathbf{b} \mid \mathbf{y}_{1:t}, \theta_t)\, d\mathbf{b}$ and inserting eqn. (19), the integral can be solved and is found to be Gaussian with mean, $\hat{\mu}_{t|t}$, and covariance, $\hat{\Sigma}_{t|t}$, where

$$\hat{\mu}_{t|t} = \alpha_{t|t} + \Gamma_{t|t}\mu_{\mathbf{b},t} \quad \text{and} \quad \hat{\Sigma}_{t|t} = \Sigma_{t|t} + \Gamma_{t|t}\Sigma_{\mathbf{b},t}\Gamma_{t|t}^T. \tag{22}$$

Recalling eqn. (19), the marginalised mean is thus equivalent to inserting the maximum *a posteriori* (MAP) estimate of the channel in the KF update in eqn. (17). The likelihood, $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}, \mathbf{b})$, is obtained by marginalising the channel from eqn. (18), i.e.,

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}) = \mathcal{N}\left(\mathbf{y}_t \,\middle|\, \mu_{y_t}, \Sigma_{\mathbf{z}_{t,\mathbf{b}}}\right). \tag{23}$$

where $\mu_{y_t} \triangleq \mathrm{Y}_{t-1}\mu_{\mathbf{b},t-1} + \mathrm{C}^T\left(\alpha_{t|t-1} + \Gamma_{t|t-1}\mu_{\mathbf{b},t-1}\right)$.

Therefore, N samples of $\theta_t^{(i)}$ are drawn from the prior importance distribution, $p\left(\theta_t^{(i)} \,\middle|\, \theta_{t-1}^{(i)}\right)$. For each particle, the channel is estimated using eqn. (20) and the linearity parameters, $\alpha_{t|t-1}$ and $\Gamma_{t|t-1}$ are evaluated. The source signal is estimated using eqn. (22). The particles are then resampled based on the likelihood in eqn. (23). The final estimate of the unknown variables at t is given by the particle average (see, e.g. [4, 5]).

5. RESULTS

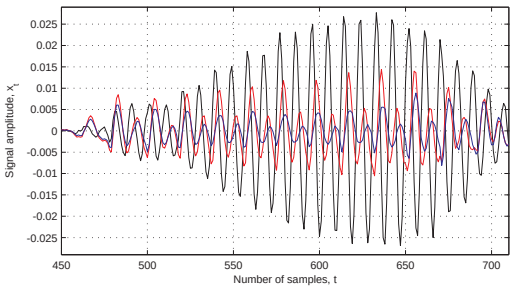
This section compares the performance of the Marginalized Rao-Blackwellized (MARBLE) particle filter using the Markov chain based and the proposed PFS speech model. To test the performance for different phoneme types, a database of ten sentences uttered by a female American speaker from the TIMIT database and recorded at $f_s = 16$kHz is segmented into the four speech sequences, containing only 1. vowels (e.g., /iy/, /ae/), 2. stop consonants (e.g., /b/, /d/), 3. fricatives (e.g., /sh/, /z/), and 4. semivowels (e.g., /r/, /l/) . The sequences are downsampled to $f_s = 4$kHz, distorted by WGN of signal-to-noise ratio (SNR) 35dB and filtered by an acoustic gramophone horn response investigated in [13]. The MARBLE particle filter is executed for 1000 particles assuming 15 TVAR parameters for the Markov chain based model and three resonators for the PFS model. The horn response can be modelled by an all-pole filter of order 8 according to [13].

The segmental SNR is evaluated for the estimated and observed signals and summarised in Table 1. For both models, the MARBLE particle filter achieves significant enhancement of the distorted signals of up to 9.75dB. The PFS model outperforms the Markov chain based model for vowels, stops, and fricatives, whereas the Markov chain based model is more appropriate for semivowels.

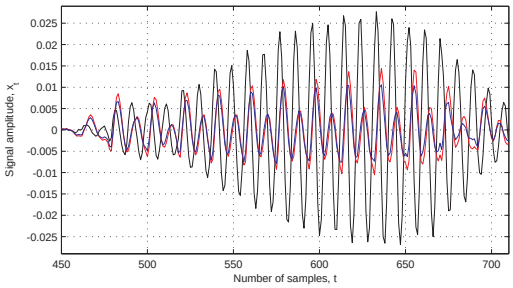
The experiment is repeated for the speech utterance 'she' at 4kHz, again reverberated by the gramophone horn and WGN. The segmental SNR of the observed signal is -4.78dB. The MARBLE particle filter achieves an improvement of 10.76dB for the Markov chain based model with an estimated SNR of 5.98dB. An improvement of 12.07dB is achieved for the PFS model with an estimated SNR of 7.29dB. The source signal for the segment 'e' in 'she' is compared to the reverberant observed signal and the estimated signal for both models in Fig. 4. Whilst the estimated signal for the Markov chain based model is slightly attenuated in amplitude as compared to the source signal, the PFS model achieves a good approximation of the variation of the speech segment.

6. CONCLUSION

This paper extended the Markov chain based source model used in the MARBLE particle filter to a novel PFS model parameterised in terms of the PARCOR coefficients. Experimental results showed that the proposed model facilitates improved speech modelling particularly for vowels, fricatives, and stop consonants, with SNR improvements of up to 9.75dB.



(a) Markov chain based source model.



(b) PFS model.

Figure 4: Comparison anechoic source signal (red) with observed signal (black) and estimated signal (blue) for the PFS and Markov chain based model of the segment 'e' in 'she'.

References

[1] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. IEEE Conf. IWAENC*, Eindhoven, Netherlands, 2005.

[2] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[3] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 6, pp. 1791–1801, Aug. 2007.

[4] C. Evers and J. R. Hopgood, "Marginalization of static observation parameters in a Rao-Blackwellized particle filter with application to sequential blind speech dereverberation," in *Proc. EUSIPCO*, Glasgow, UK, Aug. 2009.

[5] ——, "Multichannel online blind speech dereverberation with marginalization of static observation parameters in a Rao-Blackwellized particle filter," *Springer J. Signal Process. Systems*, 2009, in print.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[7] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 320–328, Apr. 1994.

[8] C. Evers, J. R. Hopgood, and J. Bell, "Acoustic models for online blind source dereverberation using sequential monte carlo methods," in *Proc. IEEE Conf. ICASSP*, Las Vegas, NV, 24 Mar. - 4 Apr. 2008.

[9] C. W. Therrien, *Discrete random signals and statistical signal processing*, ser. Signal processing series. Englewood Cliffs, NJ: Prentice Hall, 1992.

[10] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: Forecasting and control*, 3rd ed. New Jersey, NJ: Prentice Hall, 1994.

[11] T. Beierholm and O. Winther, "Particle filter inference in an articulatory-based speech model," *IEEE Sig. Process. Lett.*, vol. 14, no. 11, pp. 883–886, Nov. 2007.

[12] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter - Particle Filters for Tracking Applications*. Artech House, 2004.

[13] P. S. Spencer, "System identification with application to the restoration of archived gramophone recordings," PhD Thesis, University of Cambridge, UK, Jun. 1990.