# Towards single-channel blind dereverberation of speech from a moving speaker

**By James R. Hopgood and Christine Evers**

james.hopgood@ed.ac.uk, c.evers@ed.ac.uk
Institute for Digital Communications
Joint Research Institute for Signal and Image Processing
School of Engineering and Electronics, The University of Edinburgh

## Abstract

An overview of some key issues involved in single-channel blind dereverberation is presented, beginning with a discussion on the nature of time-varying acoustic channels, the problem of long acoustic impulse responses and an introduction to model-based solutions. The paper then discusses an approach in which the acoustic source is modelled by a time-varying AR process, and the channel by a linear time-varying all-pole filter. In each case, the time-varying filter coefficients are modelled as a linear combination of basis functions. Bayesian inference is used to estimate these coefficients. Although the propsed model needs validating against real acoustic sources and channels, it is proposed as an initial step towards solving the single-channel blind dereverberation.

## 1. Introduction

### 1.1. *Acoustic Reverberation*

Audio signals acquired in enclosed acoustic environments exhibit reverberation due to the physical separation between the source and microphone. This effectively distorts the source signal through spectral colouration reducing the intelligibility of speech. Blind acoustic dereverberation attempts to reduce the effect of this spectral distortion.

Multiple microphone blind dereverberation techniques utilise the spatial diversity of the acoustic channels. However, there are still numerous applications where effectively only a single measurement of the reverberant signal is available. *Single-channel blind dereverberation* is essential in applications where microphone arrays prove impractical, or are ineffectual due to the physical size of the array. Examples include hearing aids, hands-free telephony and automatic speech recognition where it is found more difficult to identify reverberant natural speech, rather than anechoic closely coupled speech.

### 1.2. *Time-Varying Acoustic Channels*

In some applications, the source-sensor geometry is either fixed or not varying rapidly; for example, a hands-free kit in a car cabin where the relative positions of the speaker and microphone are fixed, or in a work environment where a user is seated in front of a computer terminal. Naturally, there are many scenarios in which the source-sensor geometry is subject to change; the wearer of a hearing-aid will typically wish to move around a room, as might users of hands-free conference telephony equipment. Even in applications where the source-sensor geometry is fixed, the room acoustics may vary: the changing state of doors, windows, or moving objects will influence the room dynamics.

A speaker moving around a room at 1 m/s, equivalent to around 4 kph, covers a distance of 50 mm in 50 msec. This distance might be enough for the acoustic impulse

response to vary sufficiently significantly that any assumption of a slowly varying or time-invariant acoustic channel is no longer valid. Although there is some recent work dealing with time-varying acoustic channels [Daly *et al.*, 2004], generally the problem of single-channel blind dereverberation of speech from a moving speaker has not been addressed by the signal processing community.

## 2. Subband Methods

Single-channel blind dereverberation is a notoriously difficult and challenging problem, so much so that some researchers believe it is impossible to achieve any notable improvement in audio quality. There are two distinct approaches to the problem. The first is an optimal filtering problem in which estimates of the unknown source signal are estimated directly from the reverberant data [Daly *et al.*, 2004]. A second approach involves attempting to blindly estimate the acoustic channel from the reverberant data, followed by deconvolving the reverberant signal with the channel estimate to obtain the anechoic signal.

### 2.1. *Inverting acoustic impulse responses*

Problems encountered when dealing with acoustic channels [Radlović *et al.*, 2000] are:

(*a*) The length of acoustic impulse responses (AIRs) are of the order $P = f_s\, T_{60}$, where $T_{60}$ is the reverberation time and $f_s$ is the sampling frequency. For a room with $T_{60} = 0.5$ sec and $f_s = 10kHz$, a FIR implementation of the channel needs $P = 5000$ taps.

(*b*) AIRs are non-minimum phase and causes difficulties with inversion. The contribution of the nonminimum-phase component to the perception of reverberation is extremely important [Johansen & Rubak, 1996,Radlović & Kennedy, 2000] and must not be ignored.

(*c*) Any small error in the estimate of an AIR leads to a significant error in the inverse of the AIR. Thus, inversion can increase distortion in the enhanced signal compared to the measured reverberant signal. In particular, any deviation from the true AIR implies that attempts to equalise high-Q resonances can still leave high-Q resonances in the equalised response degrading the intelligibility of the enhanced signal.

(*d*) Similarly, while a small change in source-sensor geometry might give rise to a small change in the AIR, the corresponding changes in the inverse of an AIR are large.

### 2.2. *Subband Audio and Acoustic Processing*

Some of these problems can be alleviated by neither attempting to process the full frequency range of the source, nor attempting to invert the *full-band* room transfer function (RTF) using a single filter. In problems with long channels, it is better to utilise subband methods that attempt to enhance the reverberant signal or invert the channel response over a number of separate frequency ranges. For example, modelling each frequency band independently can lead to a parsimonious approximation of the RTF, lower model orders, and an overall reduction in the total number of parameters needed to approximate the acoustic channel [Hogood, 2005]. Moreover, there may be only a few bands that have high-Q resonances needing careful equalisation, whereas other frequency bands have lower Q factors, so less care is required.

## 3. Model-based blind dereverberation

Single-channel blind dereverberation is an inherently under-determined problem. For example, if both the source and channel are modelled as autoregressive (AR) processes, the observed signal is an AR process as well. Consequently, it is not possible to attribute a particular pole estimate to either the source or the channel, and thus there is an

ambiguity problem *if* the source signal is stationary. Source-channel ambiguities can be avoided by, for example, modelling the acoustic source as an time-varying AR (TVAR) process, and the channel by an FIR filter. In this case, the observed signal is a time-varying ARMA process, in which the poles belong to the source model and the zeros belong to the channel. Hence, there appears to be no ambiguity in distinguishing between the parameters associated with each. This model has been investigated in [Daly *et al.*, 2004] for the case of separating and recovering convolutively mixed signals. This is not always a realistic model, however, as it cannot be ascertained that the source only has poles and no zeros, and the channel only has zeros, and no poles.

### 3.1. *Block-stationary source and time-invariant channel model*

A previous approach to single-channel blind dereverberation [Hopgood & Rayner, 2003] assumes that the ensemble statistics of speech signals remain approximately stationary for around 20 – 50 msec, such that the source can be modelled by a block stationary AR (BSAR) process, and that the AIR can be modelled by a linear time-invariant (LTI) all-pole filter. The locally-stationary nature of the source signal and the time-invariance of the channel provide sufficient information to be able to distinguish between between the two models during the estimation process. This allows the acoustic channel to be uniquely identified upto a scaling ambiguity. These models yield good results when applied to relatively simple acoustic environments [Hopgood & Rayner, 2003].

*Bayesian inference* provides a rigorous and robust framework for model-based signal processing, while numerical optimisation methods provide practical algorithms. Following the procedure in [Hopgood & Rayner, 2003], the source parameters are marginalised to yield an estimate for the channel parameters, which are then used for dereverberation.

### 3.2. *Time-varying acoustic source and channel model*

To begin to address the problem of a time-varying acoustic channel, this paper extends the work in [Hopgood & Rayner, 2003] by representing the channel as a linear time-varying all-pole filter, in which the parameters of the filter are modelled as a linear combination of known basis function with unknown weightings. These basis functions might include Fourier series expansions or low-order polynomials; this model is discussed in [Rajan & Rayner, 1995]. The speech signal, rather than being a BSAR, is also a TVAR process whose parameters are represented by a linear combination of basis functions.

It is proposed that the acoustic source, $s(n)$, is modelled as a TVAR process, where:

$$s(n) = -\sum_{q=1}^{Q} b_q(n)\, s(n-q) + e(n), \qquad b_q(n) = \sum_{k=1}^{F} b_{qk}\, f_k(n-q) \tag{3.1}$$

where $e(n) \sim \mathcal{N}\left(0,\, \sigma_e^2\right)$ is the source excitation sequence, the $Q$ time-varying coefficients $\{b_q(n)\}_1^Q$ are modelled as a linear combination of $F$ *known* basis functions, $\{f_k(n)\}_1^F$, with *unknown* source coefficients, $\mathbf{b} = \{b_{qk}\}_{k=1}^F$. Similarly, the observed signal, $x(n)$, is:

$$x(n) = -\sum_{p=1}^{P} a_p(n)\, x(n-p) + s(n), \qquad a_p(n) = \sum_{k=1}^{G} a_{pk}\, g_k(n-q) \tag{3.2}$$

where $\mathbf{a} = \{a_{pk}\}_{k=1}^G$ are the *unknown* acoustic channel coefficients. [Rajan & Rayner, 1995] show equation (3.1) can be written as a linear-in-the-parameters (LITP) model. This model facilitates for formulation of the *likelihood function* and, following [Hopgood & Rayner, 2003], the model parameters $\mathbf{a}$, $\mathbf{b}$, and $\sigma_e^2$ are sampling from various conditional densities using the Gibbs sampler, and the MMSE estimate is used.

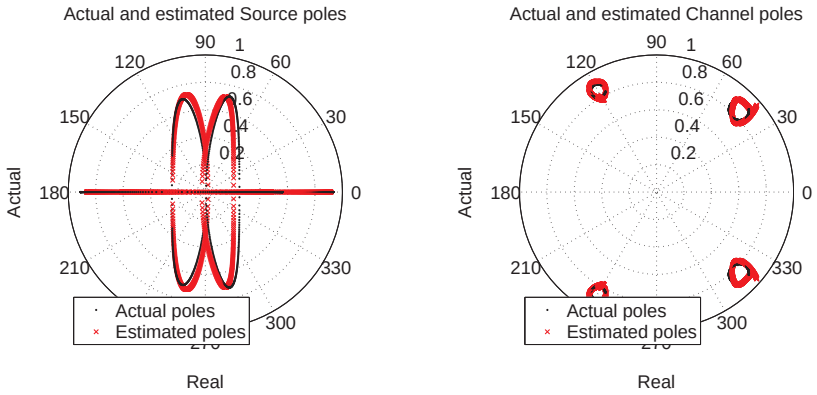


FIGURE 1. True and estimated source poles (left) and channel poles (right).

As an illustrative toy example, Figure 1 shows the trajectories of the poles of a simulated $2^{nd}$-order TVAR process which is the input of a $4^{th}$-order time-varying all-pole filter. These trajectories are chosen to give some similarity to the variation of poles seen in speech and time-varying acoustic channels. The source basis functions are $\{f_k(n)\}_1^4 = \{1, \sin(2\pi\frac{n}{N}), \cos(2\pi\frac{n}{N}), \sin(4\pi\frac{n}{N}), \cos(4\pi\frac{n}{N})\}$ where $N = 4000$ is the number of samples. Similarly, for the channel $\{g_k(n)\}_1^4 = \{1, \sin(2\pi\frac{n}{N}), \cos(2\pi\frac{n}{N}), \sin(2.5\pi\frac{n}{N}), \cos(2.5\pi\frac{n}{N})\}$. These basis functions are selected based on their ability to model the parameter variation. The estimates of the pole trajectories are the MMSE estimates from a Gibbs sampler run for 400 iterations with a burn-in of 100 iterations. Clearly this is a very simple simulated example, yet it indicates it might be possible to extend the work of [Hopgood & Rayner, 2003] to deal with time-varying channels. Further results will be presented at the conference, along with a discussion of a possible model ambiguity.

## 4. Conclusions

Model-based approaches are fundamentally based on the availability of realistic and tractable models that reflect the underlying speech processes and acoustic systems. Further work is needed to investigate whether the presented models can model real room acoustics, and thus it is necessary to analyse real AIRs in more depth. The work in [Rajan & Rayner, 1995] suggests the proposed TVAR process can model speech well.

## REFERENCES

M. Daly, J. Reilly, and J. Manton, "A bayesian approach to blind source recovery," in *Asilomar Conference on Signals, Systems, and Computers*, 2004.

B. D. Radlović, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. SAP*, vol. 8, no. 3, pp. 311–319, May 2000.

L. G. Johansen and P. Rubak, "The excess phase in loudspeaker/room transfer functions: Can it be ignored in equalization tasks?" in *J. of the AES (abstracts)*, May 1996, preprint 4181.

B. D. Radlović and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. SAP*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

J. R. Hopgood, "A subband modelling approach to the enhancement of speech captured in reverberant acoustic environments: MIMO case," in *Proc. IEEE WASPAA*, Oct. 2005.

J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. SAP*, vol. 11, no. 7, pp. 476– 488, Sept. 2003.

J. J. Rajan and P. J. W. Rayner, "Parameter estimation of time-varying autoregressive models using the Gibbs sampler," *Electronic Letters*, vol. 31, no. 13, pp. 1035–1036, June 1995.