ERPem
Edinburgh Research Partnership in Engineering and Mathematics

The University of Edinburgh
Heriot Watt University
Napier University Edinburgh

Christine Evers, James R. Hopgood *and* Judith Bell

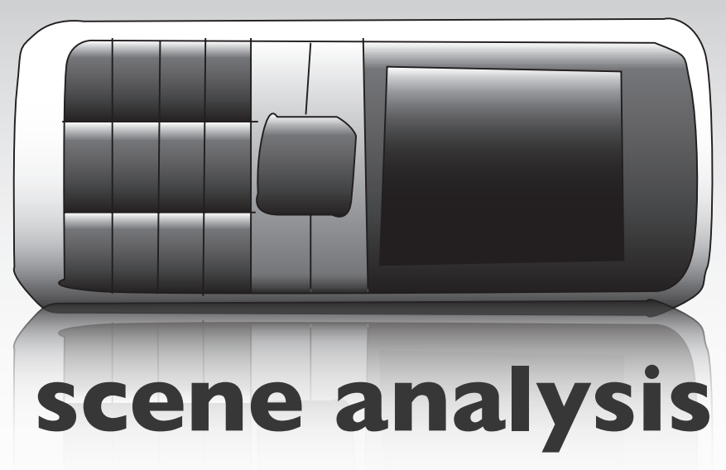
JOINT TRACKING *and* ENHANCEMENT *of* AUDIO SIGNALS *in* REVERBERANT ENVIRONMENTS

Overview

Environmental noise, music, and acoustic reverberation often lead to ***unintelligibility*** in speech in applications such as mobile phones or hearing aids. ***Enhancement*** of the degraded noisy signal is an extremely important Engineering problem for many real world audio communication systems. Enhancement relies on ***prior information*** about the speech, noise, and room acoustics. As exact prior knowledge is often not available, we ***model belief*** about the system to find the most probable restored signal.

Applications

Speech processing



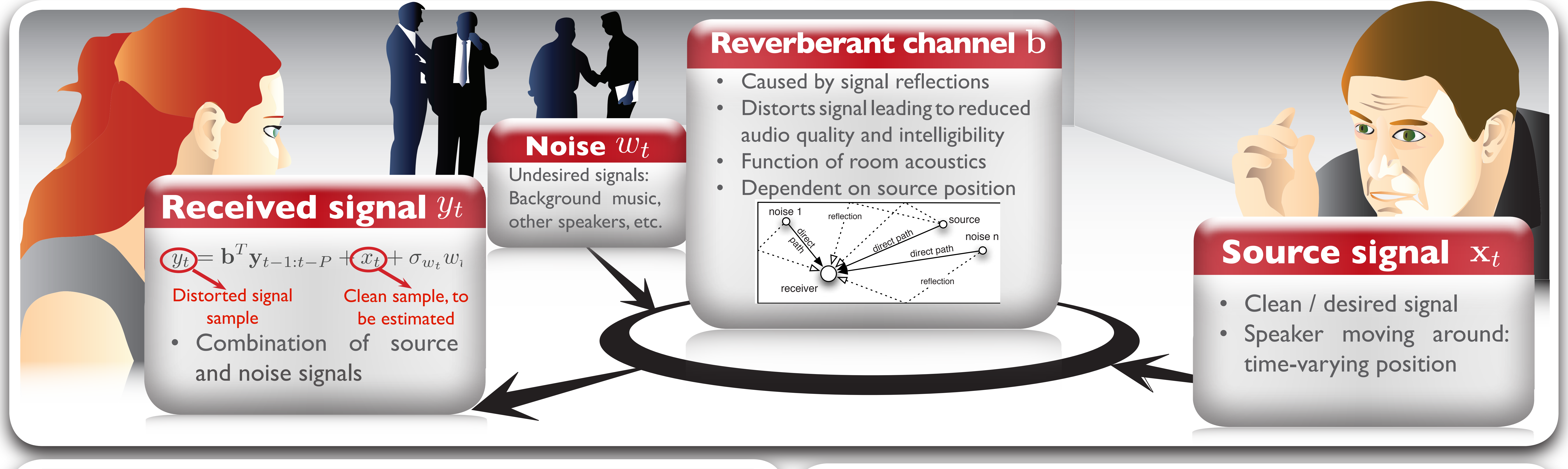
Hearing aids, teleconferencing, speech recognition systems, handsfree telephones

Underwater acoustics & scene analysis

Target localization, tracking, identification, classification, detection

Military & security applications

Object guidance, air traffic control, surveillance



Received signal y_t

$$\boxed{y_t} = \mathbf{b}^T\mathbf{y}_{t-1:t-P} + \boxed{x_t} + \sigma_{w_t}w_t$$

Distorted signal sample

Clean sample, to be estimated

- Combination of source and noise signals

Noise w_t

Undesired signals: Background music, other speakers, etc.

Reverberant channel $\mathbf{b}$

- Caused by signal reflections
- Distorts signal leading to reduced audio quality and intelligibility
- Function of room acoustics
- Dependent on source position

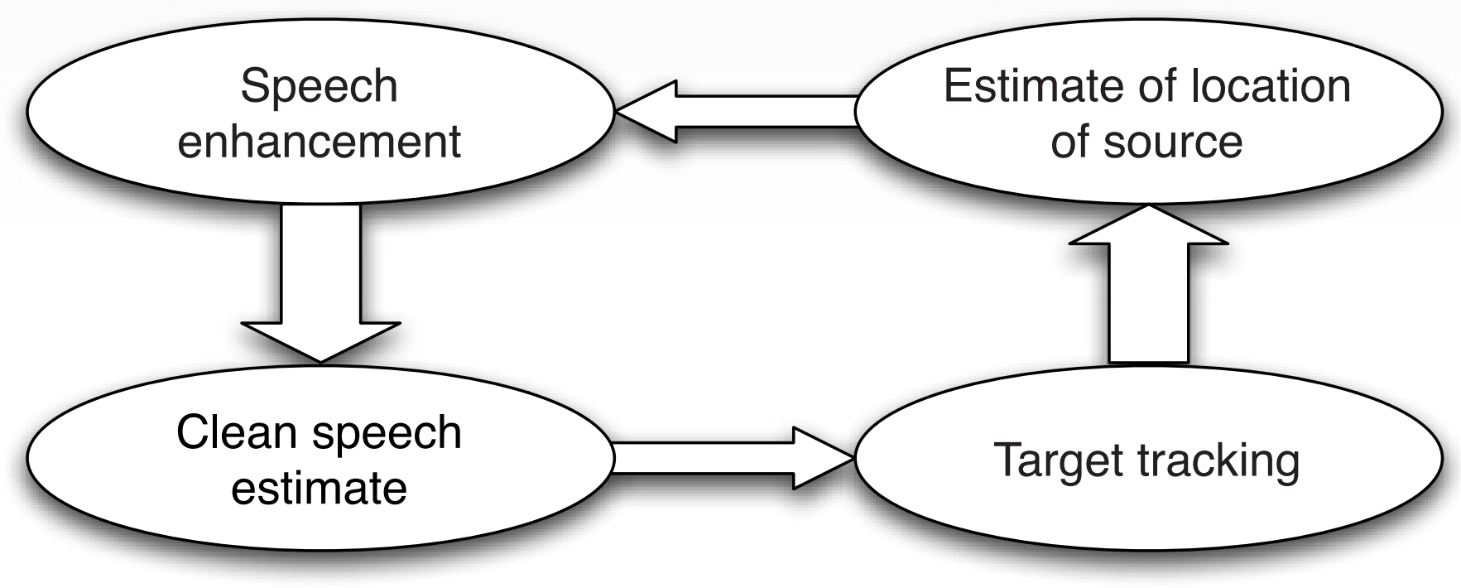
Source signal $\mathbf{x}_t$

- Clean / desired signal
- Speaker moving around: time-varying position

Problem: Joint Enhancement & Tracking

Enhancement of source signal

- Removes noise and reverberation from received signal
- Dereverberation: Estimate of the room acoustics is required
- Improve estimate of room acoustics using source location



Tracking of audio source

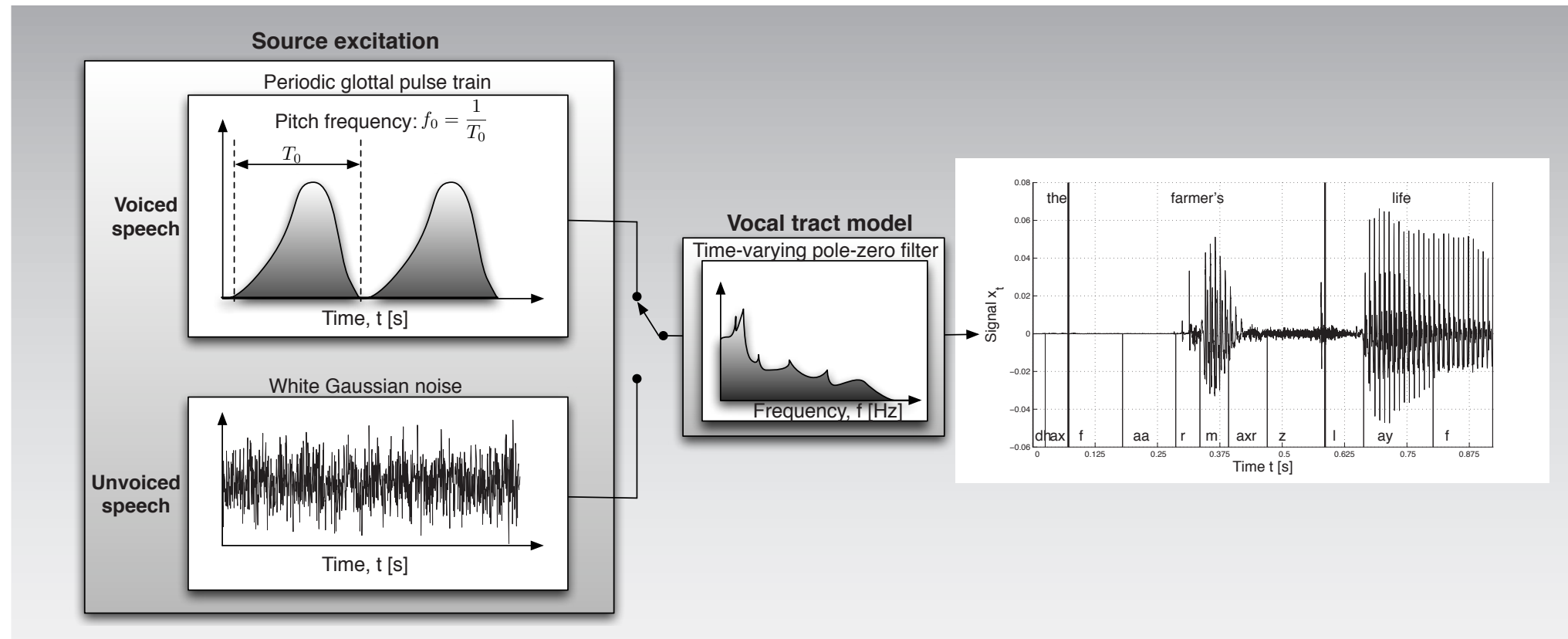
- Estimates source location
- Source signal estimate necessary to identify true source from reflections and noise

Bayesian Blind Dereverberation

Bayesian probabilistic inference allows us to introduce uncertainty in estimation methods of the unknown parameters and to update our belief in the estimates as new data becomes available.

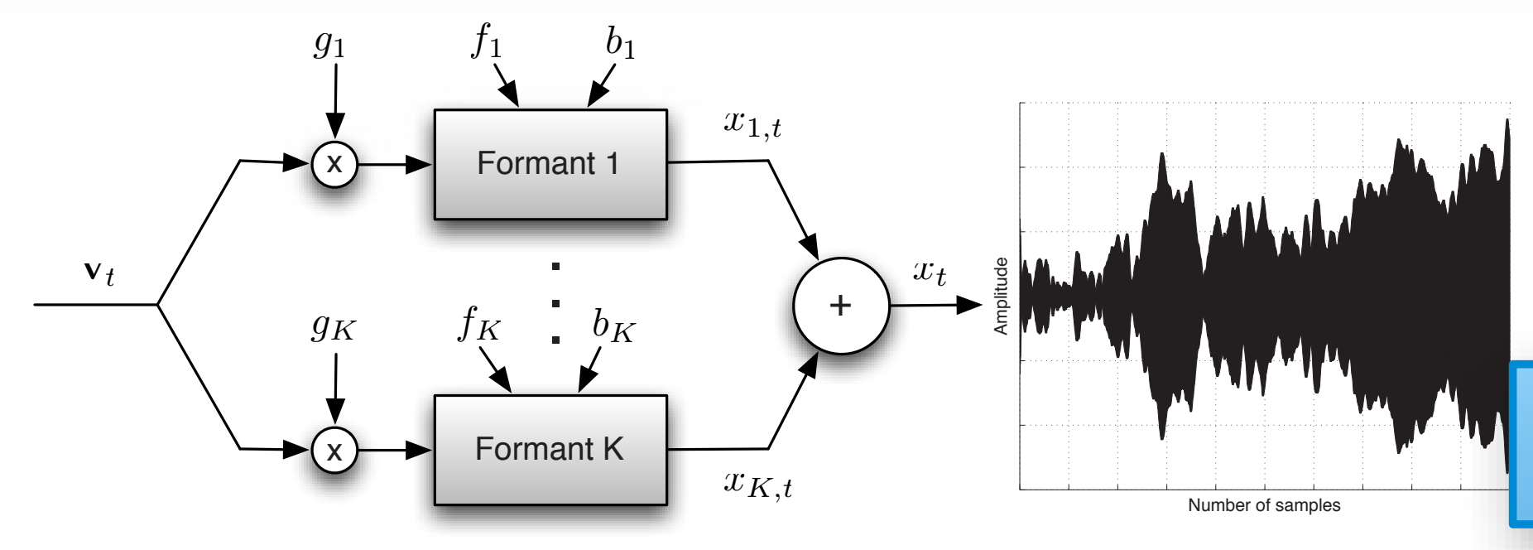
$$\underbrace{p\left(\mathbf{x}_{0:t},\boldsymbol{\theta}_{0:t}\mid\mathbf{y}_{1:t}\right)}_{\text{Joint posterior}} = \frac{\overbrace{p\left(\mathbf{y}_{1:t}\mid\mathbf{x}_{0:t},\boldsymbol{\theta}_{0:t}\right)}^{\text{Likelihood}}\,\overbrace{p\left(\mathbf{x}_{0:t},\boldsymbol{\theta}_{0:t}\right)}^{\text{Joint prior}}}{\underbrace{p\left(\mathbf{y}_{1:t}\right)}_{\text{Evidence}}}$$

Speech Models



Speech can be modeled by autoregressive (AR) processes due to their accurate capturing of the short-term spectrum of speech. Time-varying AR (TVAR) parameters are used in order to capture the continually changing properties of the vocal tract. The TVAR parameters evolve according to a first-order Markov chain to allow for a flexible representation. Voice activity detectors distinguish between voiced and unvoiced speech sequences.

For **unvoiced speech** the AR process is excited by turbulent noise, i.e., a white Gaussian noise sequence.



The vibrations of the vocal chords occuring for **voiced speech** are represented by a glottal pulse excitation of the AR process. Parallel formant synthesizers (PFSs) can be used to approximate the harmonic behavior of voiced speech. PFSs generate speech by means of several digital resonators.

Marginalization of channel parameters

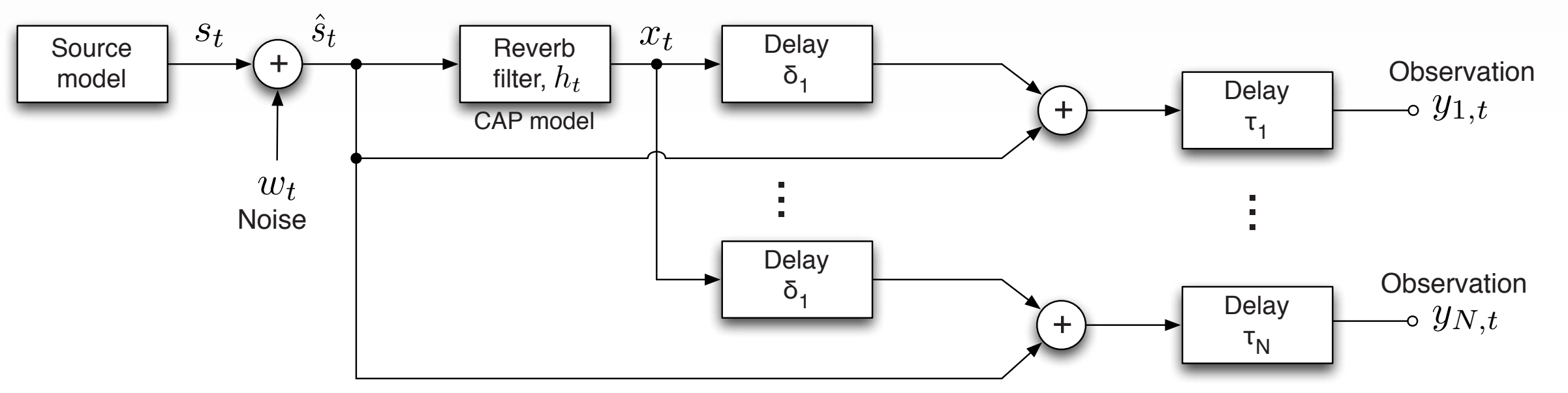
The distorting channel effects can be removed from the observed signal by ***analytically marginalizing*** the channel parameters.

$$p\left(\mathbf{x}_{0:t}\mid\mathbf{y}_{1:t},\boldsymbol{\theta}_{0:t}\right) = p\left(\mathbf{x}_0\mid\boldsymbol{\theta}_0\right)\prod_{k\in\mathcal{T}}\left[\int p\left(\mathbf{x}_t\mid\mathbf{y}_{1:t},\boldsymbol{\theta}_{0:t},\mathbf{b}\right)p\left(\mathbf{b}\mid\mathbf{y}_{1:t},\boldsymbol{\theta}_{0:t}\right)d\mathbf{b}\right]$$

As the channel parameters are usually ***unknown***, they need to be estimated like all other unknown variables. Thus, an ensemble of Kalman filters for stochastically sampled channel parameters is run simultaneously to marginalize over all the hypothesized channels.

Tracking using a time-varying channel

The source position is implicitly tracked as side-product of the channel parameter estimation by modelling the room impulse response as a function of the location of the source position relative to the sensor. The channel model is extended to the multi-sensor case.



[1] C. Evers and J. R. Hopgood, "Blind speech dereverberation using particle filters by marginalization of channel parameters", in process, *IEEE J. Trans. Speech Audio Proc.*, 2009.
[2] C. Evers and J. R. Hopgood, "Parametric modelling for single-channel blind dereverberation of speech from a moving speaker," *IET J. Sig. Proc.*, pp. 59–74, vol. 2, June, 2008.
[3] C. Evers, J. R. Hopgood, and J. Bell, "Acoustic models for blind source dereverberation using sequential Monte Carlo methods", *Proc. IEEE ICASSP*, Las Vegas, Apr. 2008
[4] C. Evers, J. R. Hopgood, and J. Bell, "Blind speech dereverberation using batch and sequential Monte Carlo methods", *Proc. IEEE ISCAS*, Seattle, May 2008
[5] J. R. Hopgood, C. Evers and S. Fortune, "Bayesian Single Channel Blind Dereverberation of Speech from a Moving Speaker", to appear in *Speech Dereverberation*, P. Naylor and N. Gaubitch (Eds.), Springer, 2008