DETECTION AND SUPPRESSION OF KEYBOARD TRANSIENT NOISE IN AUDIO STREAMS WITH AUXILIARY KEYBED MICROPHONE

*Simon Godsill, Herbert Buchner**

SigProC Laboratory, Dept.of Engineering
University of Cambridge, England

Jan Skoglund

Google Inc.
Mountain View, California, USA

ABSTRACT

In this paper a problem in transient noise suppression for audio streams in laptop and netbook devices is addressed. One or more microphones record voice signals which are corrupted with ambient noise and also transient noise from keyboard and mouse clicks. In the current work, a synchronous reference microphone is embedded in the keyboard which allows for measurement of the key click noise, substantially unaffected by the voice signal and ambient noise. An algorithm is here presented for incorporation of the keybed microphone as a reference signal in a signal restoration process for the voice part. The problem is substantially complicated by the presence of nonlinear vibarations (we postulate) in the hinge and casework of the laptop, which renders a simple linear suppressor ineffective in some cases. Moreover, the transfer functions between key clicks and voice microphone depend strongly upon which key is being clicked. A very low-latency solution is proposed in which short-time transform data is processed sequentially in short frames and a robust statistical model is formulated and estimated using Bayesian inference procedures. Results with real recordings show a significant reduction of typing artefacts at the expense of small amounts of voice distortion.

Index Terms— Key-click noise, transient noise, audio enhancement, Bayesian methods, Expectation-maximisation (EM)

1. INTRODUCTION

In many modern telephony and teleconferencing environments it is common to encounter annoying keyboard typing noise, both simultaneously present with the speech and in the 'silent' pauses between speech. A typical scenario is where someone in a conference call is taking meeting notes on their laptop or netbook while the meeting is taking place, or someone checks their emails during a voice call. Users report significant annoyance when this type of noise is present in telephony data and hence it is very desirable to remove it without introducing significant perceived distortions to the desired speech. Clearly for successful operation the processing must operate easily in real time on standard hardware and must have very low latency so that there is no irritating delay in speaker response. In more general audio restoration tasks, where real-time low-latency processing is less of an issue, model-based source separation and template-based methods have been used successfully for removing transient noise [1, 2, 3]. More modern approaches such as non-negative matrix factorisation (NMF) [4] and independent component analyis (ICA) could be possible candidates for this type of restoration, but they both have issues of latency and processing speed. As another possibility, restoration can be made more robust by inclusion of OS messages indicating which key has been pressed and when. However, the uncertain delays involved in these messages on many systems render this approach impractical at present, though likely to be of use in future systems.

So far, approaches to the keystroke removal problem have used single-ended methods in which the keyboard transients must be removed 'blind' from the audio stream without access to any timing or amplitude information about the key strikes, see for example [5, 6, 7, 8, 9, 10] for work in typing suppression and in the related area of tapping noise suppression. Clearly there are issues of reliability and signal fidelity with such approaches, and speech distortions may be audible and/or keystrokes are left untouched. In contrast with existing approaches, we introduce a reference microphone input signal for the keyboard noise, we introduce a new robust Bayesian statistical model for regressing the voice microphone on the keyboard reference microphone, which allows for direct inference about the desired voice signal while marginalising the unwanted power spectral values of the voice and keystroke noise, and we develop a simple and efficient Expectation-maximisation (EM) procedure for fast, online enhancement of the corrupted signal.

2. RECORDING SETUP

Here we study an alternative to the standard setup in which a reference microphone is available which records the sounds made by the key strikes directly, and uses this as an auxiliary audio stream to aid the restoration of the primary voice channel. We have available synchronised recordings sampled at 44.1kHz of the voice microphone waveform, X_V, and the keybed microphone waveform, X_K. The keybed microphone is placed below the keyboard in the body of the device, and is well acoustically insulated from the surrounding environment. It can be reasonably assumed to contain very little of the desired speech and ambient noise, and to serve as a good reference recording of the contaminating keystroke noise. See [11] for more detail. We will assume from now on that the audio data have been transformed into a time-frequency domain using a suitable online method such as the short-time Fourier Transform (STFT). In the case of the STFT then, $X_{V,j,t}$ and $X_{K,j,t}$ will represent complex frequency coefficients at some frequency bin j and time frame t, although we will omit these indices where it introduces no ambiguity.

3. MODELLING AND INFERENCE

A first approach might be to model the voice waveform assuming a linear transfer function H_j at frequency bin j between the reference

*We acknowledge the generous funding of Google Inc. for carrying out this work.

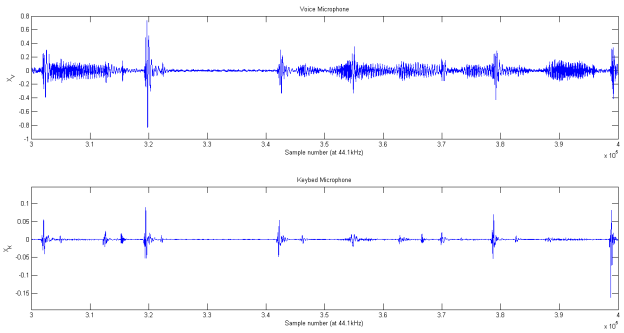

Fig. 1. Example simultaneously recorded waveforms; top: voice microphone with simultaneous speech and key strokes; bottom: Keybed microphone with (principally) just keyboard strikes.

microphone and the voice microphone, and assuming that no speech contaminates the keybed microphone: $X_{V,j} = V_j + H_j X_{K,j}$, omitting time frame index, where V is the desired voice signal and H is the transfer function from measured keybed microphone X_K to voice microphone. There are some difficult issues with this formulation, however, not least being that keystrokes from different keys will have different transfer functions, so that either a large library of transfer functions for each key needs to be learned, or the system must be very rapidly adaptive when a new key is pressed. A more serious concern, however, is that we have observed very major random differences in experimentally measured transfer functions from a real system between repeated key strikes on the same key. This we postulate is due to non-linear 'rattle'-type oscillations that are set up in typical hardware systems - in particular, the Pixel netbook that we studied has a highly nonlinear response through the hinge mechanism of the case lid. Thus, while a linear transfer function approach was found to succeed in certain examples, using for example an adaptation of the methods in [12], it was unable to completely remove the effects of the keystroke disturbances in other cases.

Here then we adopt a robust signal-based approach, in which the random perturbations and nonlinearities in the transfer function are modelled as random effects in measured keystroke waveform K at the voice microphone:

$$X_{V,j} = V_j + K_j \tag{1}$$

where V is the desired voice signal and K is the undesired keyclick.

3.1. Robust model and prior distributions

A statistical model is now formulated for both the voice and keyboard signals in the frequency domain. These models obey the known characteristics of speech signals in the time-frequency domain, i.e. sparsity and heavy-tailed (non-Gaussian) behaviour, see e.g. [13, 14, 15, 16, 17, 18, 19, 20] for our recent work in this area. Now, we model V_j as a conditional complex normal distribution with random variance that is distributed as an inverted gamma distribution, which is well known to be equivalent to modelling V_j as a heavy-tailed Student-t distribution, see [21, 22],

$$V_j|\sigma_{V_j} \sim \mathcal{N}_C(0, \sigma^2_{V_j}), \;\; \sigma^2_{V_j} \sim \mathcal{IG}(\alpha_V, \beta_V) \tag{2}$$

where '$\sim$' denotes that a random variable is drawn according to the distribution to the right, $\mathcal{N}_C$ is the complex normal distribution and $\mathcal{IG}$ is the inverted-gamma distribution [21]. The prior parameters (α_V, β_V) are tuned to match the spectral variability of speech and/or the previous estimated speech spectra from earlier frames, see Results section for more information. Such a model has been found effective in a number of audio enhancement/separation domains [13, 14, 15, 16, 17, 18], and is in contrast with other well known Gaussian or non-Gaussian statistical speech models, see e.g. [23, 24, 25, 26, 27].

A novel component of the current work is that the keyboard component K is decomposed also in terms of a heavy-tailed distribution, but with its scaling regressed on the secondary reference channel $X_{K,j}$:

$$K_j|\sigma_{K,j}, \alpha, X_{K,j} \sim \mathcal{N}_C(0, \alpha^2\sigma^2_{K,j}|X_{K,j}|^2), \;\; \sigma^2_{K,j} \sim \mathcal{IG}(\alpha_K, \beta_K) \tag{3}$$

with α a random variable which scales the whole spectrum by a random gain factor[1]:

$$\alpha^2 \sim \mathcal{IG}(\alpha_\alpha, \beta_\alpha). \tag{4}$$

We make the following *conditional independence assumption* about the prior distributions: all voice and keyboard components V and K are drawn independently across frequencies and time conditional upon their scaling parameters $\sigma_{V/K}$, and also that these scaling parameters are independently drawn from the above prior structures conditional upon the overall gain factor α. Moreover, all of these components are *a priori* independent of the value of the input regressor variable X_K. This assumption is reasonable in many cases and simplifies the form of the probability distributions considerably.

The motivation for this approach is that the frequency response between keybed microphone and voice microphone has been observed to have an approximately constant gain magnitude response across frequencies (this is modelled as the unknown gain α, but subject to random perturbations of both amplitude and phase (modelled by the IG distribution on $\sigma^2_{K,j}$)). In order to remove an obvious scaling ambiguity in the product $\alpha^2\sigma^2_{K,j}$, the maximum of the prior for $\sigma^2_{K,j}$ is set to 1. The remaining prior values are tuned to matched the observed characteristics of the real recorded datasets, see Results section.

Since the ultimate task is one of estimating V_j based only upon the observed signals X_V and X_K, a suitable object for inference will be the posterior distribution,

$$p(V|X_V,X_K) = \int_{\alpha,\sigma_K,\sigma_V} p(V, \alpha, \sigma_K, \sigma_V|X_V, X_K) d\alpha d\sigma_K d\sigma_V,$$

where (σ_K, σ_V) is the collection of scale parameters $\{\sigma_{K,j}, \sigma_{V,j}\}$ across all frequency bins j in the current time frame. From the posterior distribution we may extract the expected value $E[V|X_V, X_K]$ for a MMSE estimation scheme, or some other estimate based perhaps on a perceptual cost function, see [28, 29] for methodology. Such expectations can be handled routinely using Bayesian Monte Carlo methods, see e.g. [30, 31]. However, Monte Carlo schemes would most likely render the processing non-real-time, so we avoid these here. Instead we opt for a MAP estimation using a generalised Expectation-Maximisation (EM) algorithm:

$$\hat{V}, \hat{\alpha} = \text{argmax}_{V,\alpha} p(V, \alpha|X_V, X_K),$$

where we have included α in the optimisation because it avoids an extra numerical integration, which could be expensive.

[1] In cases where an approximate spectral *shape* is known for the scaling, say f_j, which might for example be a low-pass filter response, this can be incorporated in all the subsequent working quite simply by replacing α with αf_j throughout.

3.2. Development of EM Algorithm

In the EM algorithm latent variables to be integrated out are first defined, and these are (σ_K, σ_V) for this model. Then the algorithm operates iteratively, starting with an initial guess (V^0, α^0). At iteration i, an expectation Q of the complete data log-likelihood is computed, here[2]

$$Q((V,\alpha),(V^{(i)},\alpha^{(i)})) = E[\log(p((V,\alpha)|X_K, X_V, \sigma_V, \sigma_K))|(V^{(i)},\alpha^{(i)})]$$

where $(V^{(i)}, \alpha^{(i)})$ is the ith iteration estimate of (V, α). The expectation is taken wrt $p(\sigma_V, \sigma_K|\alpha^{(i)}, V^{(i)}, X_K, X_V)$, which simplifies under the *conditional independence assumptions* to

$$p(\sigma_V,\sigma_K|\alpha^{(i)}, V^{(i)}, X_K, X_V) = \prod_j p(\sigma_{V,j}|V_j^{(i)})\, p(\sigma_{K,j}|K^{(i)}, \alpha^{(i)}, X_{K_j}) \qquad (5)$$

where $K_j^{(i)} = X_{V,j} - V_j^{(i)}$ is the current estimate of the unwanted keystroke coefficient at frequency j.

Similarly, applying the earlier *conditional independence assumptions*, the log-conditional distribution here may be expanded over frequency bins j using Bayes' Theorem as follows:

$$\begin{aligned}\log(p((V,\alpha)|X_K, X_V, \sigma_V, \sigma_K)) &\stackrel{+}{=} \log(p(\alpha^2)) + \sum_j \log(p(V_j|\sigma_{V,j})) \\ &\quad + \log(p(X_{V,j}|X_{K,j}, V_j, \sigma_{K,j}, \alpha))\end{aligned}$$

where the notation $\stackrel{+}{=}$ means 'LHS = RHS up to an additive constant', in this case a constant that does not depend on (V, α).

The E-step thus simplifies to:

$$\begin{aligned}E[\log(p((V,\alpha)|X_K, X_V, \sigma_V, \sigma_K))|(V^{(i)},\alpha^{(i)})] &\stackrel{+}{=} E\log(p(\alpha^2)) + \sum_j E\log(p(V_j|\sigma_{V,j})) \\ &\quad + E\log(p(X_{V,j}|X_{K,j}, V_j, \sigma_{K,j}, \alpha)) \\ &= E_\alpha + \sum_j E_{V_j} + E_{K_j}\end{aligned}$$

where E_α, E_{V_j} and E_{K_j} are defined from the line above. Now, the required log-likelihood term and prior for V_j are readily obtained from Eqs. (1), (3) and (2), leading to the following expressions for the expectations E_α, E_{V_j} and E_{K_j}:

$$E_\alpha = \log(p(\alpha^2)), \;\; E_{V_j} = -\frac{1}{2}|V_j|^2 E\left[\frac{1}{\sigma_{V_j}^2}\right],$$

$$E_{K_j} = -2\log(\alpha) - \frac{|(X_{V_j} - V_j)|^2}{2\alpha^2|X_{K,j}|^2} E\left[\frac{1}{\sigma_{K,j}^2}\right].$$

Now, consider $E\left[\frac{1}{\sigma_{V_j}^2}\right]$. Under the conjugate choice of prior density as in Eq. (2), and again making use of the conditional independence assumptions as in Eq. (5)

$$\begin{aligned}&p(\sigma_{V_j}^2|V_j^{(i)}) \\ &\propto \frac{1}{2\pi\sigma_{V_j}^2}\exp\left(-\frac{1}{2\sigma_{V_j}^2}|V_j^{(i)}|^2\right)\mathcal{IG}(\sigma_{V,j}^2|\alpha_V, \beta_V) \\ &= \mathcal{IG}\left(\sigma_{V_j}^2|\alpha_V + 1, \beta_V + \frac{|V_j^{(i)}|^2}{2}\right)\end{aligned}$$

Hence, at the ith iteration:

$$E\left[\frac{1}{\sigma_{V_j}^2}\right] = \frac{\alpha_V + 1}{\beta_V + \frac{|V_j^{(i)}|^2}{2}},$$

which is the mean of the corresponding gamma distribution for $1/\sigma_{V,j}^2$. For prior mixing distributions other than the simplest inverted-gamma, this expectation could be computed numerically and stored for example in a look-up table.

By similar reasoning the conditional distribution for $\sigma_{K,j}^2$ in Eq. (5) is obtained as:

$$\begin{aligned}p(\sigma_{K_j}^2|X_{K,j}, \alpha^{(i)}, K_j^{(i)}) &\propto \frac{1}{2\pi\sigma_{K,j}^2\alpha^{i\,2}|X_{K,j}|^2} \\ &\quad \exp\left(-\frac{1}{2\sigma_{K,j}^2\alpha^2|X_{K,j}|^2}|K_j{}^{(i)}|^2\right)\mathcal{IG}(\sigma_{K,j}^2|\alpha_K, \beta_K) \\ &= \mathcal{IG}\left(\alpha_K + 1, \beta_K + \frac{|K_j^{(i)}|^2}{2\alpha^{(i)\,2}|X_{K,j}|^2}\right).\end{aligned}$$

Hence, at the ith iteration:

$$E\left[\frac{1}{\sigma_{K,j}^2}\right] = \frac{\alpha_K + 1}{\beta_K + \frac{|K_j^{(i)}|^2}{2\alpha^{(i)\,2}|X_{K,j}|^2}}.$$

Finally, having computed these expectations and substituted them into Q, the M-step would ideally maximise Q jointly wrt (V, α). Because of the complex structure of the model, this cannot be done quite in closed form for this Q function. We propose instead to take advantage of iterative formulae for maximising V with α fixed, then maximising α with V fixed at the new value, and running this for a few steps within each EM iterstion. Such a scheme is a Generalised EM, which still guarantees that the posterior probability is non-decreasing at each iteration and hence can be expected to converge to the true MAP solution with increasing iteration number.

Omitting the fairly straightforward algebraic steps in finding the maxima of Q wrt V and α, we arrive at the following M-step updates. Notation is such that we initialise the Generalised M-step at each iteration with $V_j^{(i+1)} = V_j^{(i)}$, $K_j^{(i+1)} = X_{V,j} - V_j^{(i)}$, and $\alpha^{(i+1)} = \alpha^{(i)}$, the final values from the previous iteration, and we then iterate several steps of the following fixed point equations, which refine the estimates at the new iteration $i + 1$. Note that the update for V_j is essentially a Wiener filter gain, which is applied independently and in parallel for all frequencies $j = 1, ..., , J$,

$$V_j^{(i+1)} = \frac{E\left[\frac{1}{\sigma_{V_j}^2}\right]}{E\left[\frac{1}{\sigma_{V_j}^2}\right] + \frac{E\left[\frac{1}{\sigma_{K,j}^2}\right]}{\alpha^{(i+1)\,2}|X_{K,j}|^2}} X_{V,j} \qquad (6)$$

[2]Note that this is the Bayesian formulation of EM in which a prior distribution is included for the unknowns V and α, see e.g. [32]

and for α:

$$\alpha^{(i+1)} = \sqrt{\frac{\beta_\alpha + \sum_j E\left[\frac{1}{\sigma^2_{K,j}}\right] \frac{1}{2|X_{K,j}|^2}\left(|K_j^{(i+1)}|^2\right)}{\alpha_\alpha + 1 + J}} \quad (7)$$

where J is the total number of frequency bins.

Once the EM procedure has run for a number of iterations and is satisfactorily converged, the resulting spectral components V_j are taken back to the time domain via the inverse FFT (in the STFT case) and reconstructed into a continuous signal by windowed overlap-add procesdures.

4. EXPERIMENTAL RESULTS

The methods were tested on files recorded from the Pixel Chromebook [33], sampling synchronously at 44.1kHz from the voice and keybed microphones, and processing using EM was carried out in Matlab. Frame lengths of 1024 samples were used for an STFT transform, with 50% overlap and Hanning analysis windows. It was possible to record extracts of voice alone, and then of key strokes alone, then adding together the signals recorded in order to obtain corrupted microphone signals for which 'ground truth' restorations are available. Prior parameters for the Bayesian model were fixed for the simulations as follows:

- Prior $\sigma^2_{V,j} \sim \mathcal{IG}(\alpha_V, \beta_{V,j})$ (note we now make the scale parameter β_V explicitly frequency-dependent). The degrees of freedom were fixed to $\alpha_V = 4$ in order to allow a degree of flexibility and heavy-tailed behaviour in the voice signal. The parameter $\beta_{V,j}$ was set in a frequency-dependent manner as follows: the final EM-estimated voice signal from the previous frame, $|\hat{V}_j|^2$ was used to give a prior estimate of $\sigma^2_{V,j}$ for the current frame. $\beta_{V,j}$ was then fixed such that the *mode* of the IG distribution was equal to $|\hat{V}_j|^2$, i.e. we set $\beta_{V,j} = |\hat{V}_j|^2(\alpha_V + 1)$. This encourages some spectral continuity from previous frames, which reduces artifacts in the processed audio, and also allows for some reconstruction of very heavily corrupted frames based on what has gone before.
- Prior $\sigma^2_{K,j} \sim \mathcal{IG}(\alpha_K, \beta_K)$. This is fixed across all frequencies to $\alpha_K = 3$, $\beta_K = 3$, leading to a mode at $\sigma^2_{K,j} = 0.75$.
- Prior $\alpha \sim \mathcal{IG}(\alpha_\alpha, \beta_\alpha)$: $\alpha_\alpha = 4$, $\beta_\alpha = 100,000(\alpha_\alpha + 1)$, which places the prior mode for α^2 at $100,000$, which is tuned by hand from experimental analysis of data recorded with just keystroke noise present.

Various configurations were tested for the EM, and it was found that results converged with little further improvement after around 10 iterations, with 2 sub-iterations of the generalised M-step of Eqs. (6) and (7) per full EM iteration, and these parameters were then fixed for all subsequent simulations.

One further important detail is that a time-domain detector was devised to flag corrupted frames, and processing was only applied to frames for which detection was flagged, hence avoiding unnecessary signal distortions and wasted computations through processing in uncorrupted frames. This detector comprised a rule-based combnination of detections from the keybed microphone signal and the two available (stereo) voice microphones. Within each stream, detections are based on an autoregressive (AR) error signal, much as in [1] Ch.4, and frames are flagged as corrupted when the maximum error magnitude exceeds a certain factor of the median error magnitude for that frame; full details of this and performance metrics will be presented in future publications, but the method gave near 100% correct detections in the examples tried so far.

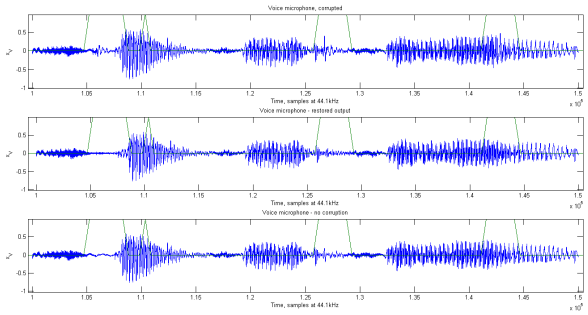

Fig. 2. Extract of typical restoration and ground truth (no keyclicks) - male speech

We evaluate performance using an average segmental SNR measure, seg-SNR $= \frac{1}{N}\sum_{n=1}^{N} 10\log_{10}\frac{\sum_{t=1}^{T} v_{t,n}^2}{\sum_{t=1}^{T}(v_{t,n}^2 - \hat{v}_{t,n})^2}$, where $v_{t,n}$ is the true, uncorrupted, voice signal at the tth sample of the nth frame, and $\hat{v}$ is the corresponding estimate of v. Performance is compared against a very simple procedure which mutes the spectral components to zero in frames which are detected as corrupted. Results show an improvement on average of approximately 3dB when taken over the whole speech extract, and of 6-10dB when including just the frames detected as corrupted. This figures may be adjusted by tuning the prior parameters to trade off perceived signal distortion against suppression levels of the noise. Although these figures may seem relatively small improvements, the perceptual effect of the EM approach is significantly improved compared with muting and with the corrupted input audio.

An example detection and restoration is shown in Fig. 4. In all three panels the frames detected as corrupted are indicated by the zero-one waveform overlaid in green. These detections agree well with a visual study of the keyclick data waveform. In the top panel we have the corrupted input voice microphone, in the middle panel the restored output, and at the bottom the original voice signal (available in this test as 'ground-truth'). Notice that the central panel manages to preserve the speech envelope and speech events around 125 ksamples and 140 ksamples, while suppressing the disturbance well around 105 ksamples. The audio is significantly improved in the restoration, leaving just a little 'click' residue which can be removed by post-processing using standard techniques [1], Ch. 4, while the simple 'muting' restoration is far too extensive to be acceptable. In this fairly typical example a favourable 10.1dB improvement in segmental SNR is obtained for corrupted frames, compared to the muting restoration, and 2.5dB improvement when all frames are considered (including the uncorrupted frames). See [11] for more detailed results and audio from these experiments.

5. CONCLUSION

We have presented new methods for enhancement of speech when corrupted by keyboard typing noise and demonstrated good performance on real data recorded from the Pixel Chromebook. However, further recent experiments have indicated that the modelling of cross-talk from voice into the keybed mic, here considered as negligible, is an area to pursue - hence future developments will include cancellation of this cross-talk into the framework, a constrained source separation framework.

6. REFERENCES

[1] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model-Based Approach*, Berlin: Springer, ISBN 3 540 76222 1, Sept. 1998.

[2] S. J. Godsill and C. H. Tan, "Removal of low frequency transient noise from old recordings using model-based signal separation techniques," in *Proc. IEEE Workshop on Audio and Acoustics, Mohonk, NY State*, Mohonk, NY State, Oct. 1997.

[3] Saeed V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, 2006.

[4] N. Mohammadiha and S. Doclo, "Transient noise reduction using nonnegative matrix factorization," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014.

[5] B. Raj, M.L. Seltzer, and R.M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[6] A. Subramanya, M.L. Seltzer, and A. Acero, "Automatic removal of typed keystrokes from speech signals," *IEEE SP Letters*, vol. 14, no. 5, pp. 363–366, May 2007.

[7] A. Sugiyama, "Single-channel impact-noise suppression with no auxiliary information for its detection," in *Proc. IEEE Workshop on Audio and Acoustics, Mohonk, NY State*, Oct. 2007.

[8] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE Workshop on Audio and Acoustics, Mohonk, NY State*, Oct. 2013.

[9] R. Talmon, I. Cohen, S. Gannot, and R.R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2528–2538, Nov 2012.

[10] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 132–144, Jan 2013.

[11] S. J. Godsill, H. Buchner, and J. Skoglund, "Website with further results from paper," http://www-sigproc.eng.cam.ac.uk/Main/SJG/ICASSP15.

[12] K. Helwani, H. Buchner, J. Benesty, and J. Chen, "Multichannel acoustic echo suppression," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 600–604.

[13] P. J. Wolfe, S. J. Godsill, and W.J. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation," *Journal of the Royal Statistical Society, Series B*, vol. 66, no. 3, pp. 575–589, 2004, Read paper (with discussion).

[14] C. Févotte and S. Godsill, "Sparse linear regression in unions of bases via Bayesian variable selection," *IEEE Signal Processing Letters*, vol. 13, no. 7, pp. 441–444, July 2006.

[15] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, 2008.

[16] C. Févotte and S.J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. on Speech and Audio Processing*, 2006.

[17] A. T. Cemgil, S. J. Godsill, and C. Fevotte, "Variational and Stochastic Inference for Bayesian Source Separation," *Digital Signal Processing*, vol. 17, 2007.

[18] S. Godsill, "The shifted inverse-gamma model for noise-floor estimation in archived audio recordings," *Signal Processing*, vol. 90, pp. 991–999, 2010.

[19] J. Murphy and S. J. Godsill, "Joint Bayesian removal of impulse and background noise," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2011, pp. 261–264.

[20] James Murphy and Simon J. Godsill, "Structured sparse bayesian modelling for audio restoration," in *Compressed Sensing & Sparse Filtering*, Avishy Y. Carmi, Lyudmila Mihaylova, and Simon J. Godsill, Eds., Signals and Communication Technology, pp. 423–453. Springer Berlin Heidelberg, 2014.

[21] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*, Wiley, 1972.

[22] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, 1994.

[23] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983, pp. 1118–1121.

[24] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[25] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Appl. Sig. Processing*, vol. 10, no. 1, pp. 1043–1051, 2003.

[26] R. Martin, "Speech enhancement based on minimum mean square error estimation and supergaussian priors," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, 2005.

[27] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, May 2005.

[28] P. J. Wolfe and S. J. Godsill, "Perceptually motivated approaches to music restoration," *Journal of New Music Research*, vol. 30, no. 1, pp. 83–92, 2001, Special issue: Music and Mathematics.

[29] S. J. Godsill, P. J. Wolfe, and W. N. W. Fong, "Statistical model-based approaches to audio restoration and analysis," *Journal of New Music Research*, vol. 30, no. 4, pp. 323–338, 2001, Special Issue: Conservation, Restoration and Archiving of Electroacoustic Music.

[30] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, 2nd edition, 2004.

[31] J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2001.

[32] M. A. Tanner, *Tools for Statistical Inference, Second Edition*, Springer-Verlag, 1993.

[33] The Chromium projects website, "Chromebook pixel," http://www.chromium.org/chromium-os/developer-information-for-chrome-os-devices/chromebook-pixel.