AN ANALYSIS OF THE EFFECT OF LARYNX-SYNCHRONOUS AVERAGING ON DEREVERBERATION OF VOICED SPEECH

Alastair H. Moore, Patrick A. Naylor

Imperial College London
Department of Electrical and Electronic Engineering

Jan Skoglund

Google Inc.
Mountain View, CA 94043

ABSTRACT

The SMERSH algorithm is a physiologically-motivated approach to low-complexity speech dereverberation. It employs multichannel linear prediction to obtain a reverberant residual signal and subsequent larynx-synchronous temporal averaging to attenuate the reverberation during voiced speech. Experimental results suggest the method is successful but, to date, no detailed analysis of the theoretical basis of the larynx-synchronous averaging has been undertaken. In this paper the SMERSH algorithm is reviewed before focussing on the theoretical basis of its approach. We show that the amount of dereverberation that can be achieved depends on the coherence of reverberation between frames. Simulations show that the extent of dereverberation increases with reverberation time and give an insight into the tradeoff between dereverberation and speech distortion.

Index Terms— dereverberation, linear prediction

1. INTRODUCTION

Dereverberation of speech has applications in automatic speech recognition and hands-free telecommunications [1]. The later is especially challenging as systems must operate with imperceptible delay and be robust to changes in the multipath propagation caused by movement of the speaker and/or microphone.

Dereverberation algorithms which operate on the linear prediction (LP) residual [2] use knowledge of the structure of speech (or the physiology of the speech production system) to increase the signal-to-reverberation ratio (see [3] for a review). One such algorithm is the Spatiotemporal averaging Method for Enhancement of Reverberant Speech (SMERSH) [4] in which the periodic nature of voiced speech is exploited. The LP residual of clean speech is characterised by an impulsive feature (epoch) associated with each glottal closure. The LP residual of reverberant speech retains these strong peaks but is contaminated by 'spurious' peaks and noise-like disturbance. By performing larynx-synchronous intercycle averaging between successive epochs the features which are common across successive cycles are enhanced.

In this contribution we present an analysis of the theoretical basis for intercycle averaging and consider the trade-offs involved. Specifically we

- derive a model for the reverberant residual as the sum of contributions from each larynx cycle (Sec. 3)
- demonstrate that the attenuation of reverberation is dependent on the coherence of the averaging, which is itself dependent on the reverberant tail (Sec. 4.2)
- demonstrate the trade-off between dereverberation and speech distortion as a function of the number of cycles in the average and of the windowing function used (Sec. 4.3).

To set the work in context we first review the SMERSH algorithm.

2. REVIEW OF SMERSH

The SMERSH algorithm [4][5] contains eight distinct operations. 1) Time-alignment of the reverberant speech signals from multiple microphones using GCC-PHAT [6] to estimate the time differences of arrival. 2) Estimation of the clean speech LP coefficients from the time aligned microphone signals using multichannel linear prediction [7]. 3) Calculation of the (spatially) averaged reverberant residual by inverse filtering the time-aligned microphone signals with the LP filter and summing across channels. 4) Estimation of the epochs from the time-aligned microphone signals using multichannel DYPSA [8]. 5) Intercycle averaging of the spatially averaged reverberant residual. 6) Estimation of an 'equivalent' linear filter which transforms the reverberant residual to the temporally averaged reverberant residual. 7) Leaky integration of the equivalent filters to produce a slowly time-varying filter which is updated on each cycle of voiced speech and applied to the spatially (but not temporally) averaged residual, thus allowing the unvoiced speech to be filtered as well. 8) Resynthesis of speech using the LP filter and the dereverberated residual signal.

The fifth process is predicated on the hypothesis that the effect of reverberation is seen in the LP residual as energy which is uncorrelated with the clean residual and so can be attenuated by averaging across neighbouring cycles. In Sec. 3

we develop a signal model for the reverberant residual which allows this assumption to be explored in more depth.

The SMERSH algorithm exploits spatial diversity to estimate the epochs and to estimate the clean speech LP coefficients but the temporal averaging operates on the (single channel) output of a spatial beamformer. Therefore, for clarity of presentation, this paper focusses on the effect of larynx-synchronous averaging of a single channel signal with epochs and LP coefficients being estimated directly from clean speech.

3. THEORETICAL BASIS OF INTERCYCLE AVERAGING

3.1. Signal model

We express a speech signal, $s(n)$, where n is the discrete time sample index, as the summation of non-overlapping frames according to

$$s_i(n) = \begin{cases} s(n) & n = \{\kappa_i, \kappa_i + 1, \ldots, \kappa_i + L_i - 1\} \\ 0 & \text{otherwise} \end{cases}$$

where the i-th frame is defined by its starting sample index, κ_i, and length, L_i, such that $s(n) = \sum_{i=0}^{\infty} s_i(n)$. The noise-free observation, $x(n)$, of $s(n)$ some distance from the speaker in a reverberant room is determined by the L_h-tap impulse response, $h(n)$, according to $x(n) = h(n) * s(n)$. For the purposes of the following analysis $h(n)$ is assumed to be constant. Thus

$$x(n) = h(n) * \sum_{i=0}^{\infty} s_i(n) = \sum_{i=0}^{\infty} h(n) * s_i(n) \quad (1)$$

and taking the z-transform gives

$$X(z) = \sum_{i=0}^{\infty} H(z) S_i(z). \quad (2)$$

3.2. Linear prediction decomposition

Using linear prediction, $S_i(z)$ can be decomposed into an excitation signal, $E_i(z)$, driving an all-pole filter, $1/A_i(z)$, such that

$$X(z) = \sum_{i=0}^{\infty} H(z) \frac{1}{A_i(z)} E_i(z). \quad (3)$$

Linear prediction analysis of $x(n)$ over the interval $n = \{\kappa_j, \kappa_j + 1, \ldots, \kappa_j + L_j - 1\}$ yields the all-pole filter $1/\hat{A}_j(z)$. Applying the inverse filter, $\hat{A}_j(z)$, to $X(z)$ gives

$$\varepsilon_j(z) = \sum_{i=0}^{\infty} H(z) \frac{\hat{A}_j(z)}{A_i(z)} E_i(z) = \sum_{i=0}^{\infty} \varepsilon_{j|i}(z) \quad (4)$$

where $\varepsilon_{j|i}(z)$ is the contribution to $\varepsilon_j(z)$ due to $E_i(z)$.

It was shown in [7] that a good approximation to the all-pole filter $1/A_j(z)$ associated with clean speech can be obtained from reverberant speech over the same time interval by exploiting spatial expectation over multiple microphones. We therefore assume that $1/\hat{A}_j(z) \approx 1/A_j(z)$ such that $\varepsilon_j(z)$ can be expressed as

$$\varepsilon_j(z) \approx H(z) E_j(z) + \sum_{i=0}^{\infty, i \neq j} \varepsilon_{j|i}(z). \quad (5)$$

The first term, $\varepsilon_{j|j}(z) \approx H(z) E_j(z)$, or equivalently in the discrete time domain, $\epsilon_{j|j}(n) \approx h(n) * e_j(n)$, is the convolution of the clean excitation signal for the j-th frame with the channel. Therefore the contribution of $\varepsilon_{j|j}(z)$ to $\varepsilon_j(z)$ is non-zero for $\kappa_j \leq n \leq (\kappa_j + L_j - 1) + L_h - 1$. The individual terms in the summation $\varepsilon_{j|i}(z), i \neq j$ can be interpreted as the convolution of the clean excitation signal for the i-th larynx cycle with the channel and an additional filtering process, $\hat{A}_j(z)/A_i(z)$, due to the mismatch between the autoregressive filters estimated from the i-th frame of clean speech and the j-th frame of reverberant speech. Thus

$$\epsilon_j(n) = \epsilon_{j|j}(n) + \sum_{i=0}^{\infty, i \neq j} \epsilon_{j|i}(n) \quad (6)$$

where $\epsilon_{j|i}(n)$ is the inverse z-transform of $\varepsilon_{j|i}(z)$.

To obtain the *reverberant residual,* $\epsilon(n)$, we apply a window

$$w_j(n) = \begin{cases} 1 & \kappa_j \leq n < \kappa_j + L_j \\ 0 & \text{otherwise} \end{cases}$$

to $\epsilon_j(n)$ and sum across all j

$$\epsilon(n) = \sum_{j=0}^{\infty} w_j(n) \epsilon_j(n). \quad (7)$$

Combining (6) and (7) and rearranging the order of summation leads to

$$\epsilon(n) = \sum_{j=0}^{\infty} w_j(n) \sum_{i=0}^{\infty} \epsilon_{j|i}(n) = \sum_{i=0}^{\infty} \epsilon_{|i}(n) \quad (8)$$

where $\epsilon_{|i}(n)$ is the contribution to $\epsilon(n)$ due to $e_i(n)$.

Linear prediction decomposition of clean, voiced speech produces a pseudo-periodic impulsive excitation signal synchronised with the glottal closures. Aligning the start of each frame to these impulsive events so as to perform larynx-synchronous processing, one can take advantage of the periodicity and the impulsive excitation. Fig. 1 shows an illustrative example of a clean speech residual signal, $e(n)$, the response of the room to three individual larynx cycles, $e_{i-1}(n)$, $e_i(n)$ and $e_{i+1}(n)$ given by $\epsilon_{|i-1}(n)$, $\epsilon_{|i}(n)$ and $\epsilon_{|i+1}(n)$, respectively, and the summation over all i to give the observed reverberant residual $\epsilon(n)$.

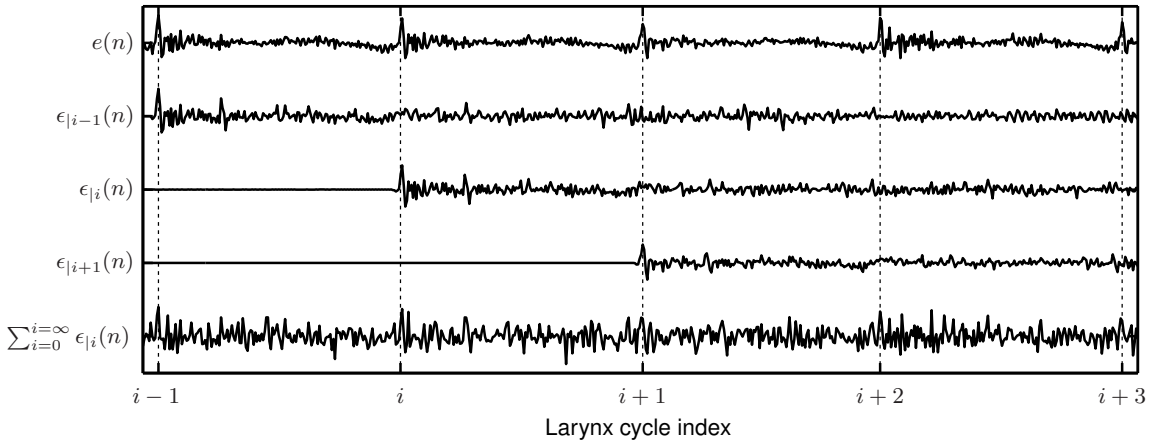

Fig. 1. Reverberant LP residual shown as sum of room response to individual glottal pulses along with clean speech LP residual

From (6) and (7) the j-th frame of $\epsilon(n)$ is given by

$$w_j(n)\epsilon_j(n) = w_j(n)\epsilon_{j|j}(n) + \sum_{i=0}^{\infty,\, i\neq j} w_j(n)\epsilon_{j|i}(n) \qquad (9)$$

which is the summation of windowed contributions from all frames of $e(n)$. The contribution from the j-th frame, $w_j(n)\epsilon_{j|j}(n)$, is the first L_j samples of the convolution $h(n) * e_j(n)$. Since the propagation time of the direct path in $h(n)$ is effectively zero (glottal closure instants are estimated from the microphone signal(s)) and the truncation is less than 20 ms (for $f_0 > 50$ Hz), only the very earliest reflections (if any) can be present in the truncated signal. We therefore make the approximation $w_j(n)\epsilon_{j|j}(n) = \tilde{e}_j(n) \approx e_j(n)$, which is the clean speech excitation signal of the j-th larynx cycle which dereverberation seeks to enhance. The reverberation, which we wish to attenuate, is the contribution from the other larynx cycles, $v_j(n) = \sum_{i=0}^{\infty,\, i\neq j} w_j(n)\epsilon_{j|i}(n)$.

3.3. Larynx-synchronous averaging

Let $\boldsymbol{\epsilon}_j = [\epsilon(\kappa_j), \epsilon(\kappa_j+1), \ldots, \epsilon(\kappa_j+L_j-1)]^T$ be the non-zero samples of the j-th frame of $\epsilon(n)$ as defined in (9) such that $\boldsymbol{\epsilon}_j = \tilde{\boldsymbol{e}}_j + \boldsymbol{v}_j$. Larynx-synchronous averaging with respect to $\boldsymbol{\epsilon}_j$ follows as

$$\bar{\boldsymbol{\epsilon}}_j \;=\; \frac{1}{c^- + c^+ + 1} \sum_{m=-c^-}^{c^+} \boldsymbol{\epsilon}_{j+m} \qquad (10)$$

where $N{=}c^-{+}c^+{+}1$ is the number of larynx cycles included in the average and each $\boldsymbol{\epsilon}_{j+m}$ is truncated or zero-padded to L_j samples to allow for cycles of slightly different length.

For voiced speech, the clean speech excitation is pseudo-periodic, so for sufficiently small c^- and c^+

$$\tilde{\boldsymbol{e}}_{j-c^-} \approx \tilde{\boldsymbol{e}}_j \approx \tilde{\boldsymbol{e}}_{j+c^+} \qquad (11)$$

with the error in the approximation depending on how quickly $\tilde{\boldsymbol{e}}_j$ changes from one larynx cycle to the next. Combining (11) with (10) gives

$$\bar{\boldsymbol{\epsilon}}_j \;= \tilde{\boldsymbol{e}}_j + \;\frac{1}{N} \sum_{m=-c^-}^{c^+} \boldsymbol{v}_{j+m}. \qquad (12)$$

Comparing (6) and (12) it is clear that larynx-synchronous averaging of the windowed reverberant residual signal has the potential to isolate the clean speech excitation signal for the target larynx cycle whilst attenuating the undesired (reverberant) contributions from surrounding larynx-cycles. The extent to which these unwanted contributions are attenuated depends on the coherence of their summation across m. This will be examined with some examples in Sec. 4.

A window function $\mathbf{w}$ of length L_j can be incorporated into (10) as

$$\bar{\boldsymbol{\epsilon}}_j \;= (\mathbf{1} - \mathbf{w}) \odot \boldsymbol{\epsilon}_j + \;\frac{1}{N} \sum_{m=-c^-}^{c^+} \mathbf{w} \odot \boldsymbol{\epsilon}_{j+m} \qquad (13)$$

where $\odot$ is the Hadamard product representing element-wise multiplication, $\mathbf{1}$ is the vector of ones of same dimensions as $\mathbf{w}$ and the rectangular window $\mathbf{w}_R = \mathbf{1}$ causes (13) to reduce to (10). In [4] a Tukey window [9]

$$\mathbf{w}_T(u) = \begin{cases} 0.5 + 0.5\cos\left(\frac{2\pi u}{\beta(L_j-1)} - \pi\right) & u < \frac{\beta L_j}{2} \\ 0.5 + 0.5\cos\left(\frac{2\pi}{\beta} + \frac{2\pi u}{\beta(L_j-1)} - \pi\right) & u > L - \frac{\beta L_j}{2} - 1 \\ 1.0 & \text{otherwise} \end{cases}$$

with taper ratio $\beta = 0.3$ was proposed. This prevents the impulsive feature associated with the glottal closure from being included in, and potentially distorted by, the averaging process. The authors suggest this is important for speech quality,

but it also limits the potential for dereverberation during this part of the cycle.

4. EXPERIMENTAL INVESTIGATION

4.1. Method

In the following experiments reverberant speech was generated by convolving a clean sample of the phoneme /ɜ/ with room impulse responses (RIRs) calculated using the image method [10]. The phoneme was sustained for 5 s by a male talker with a mean fundamental frequency 104 Hz. Using synthesised RIRs allows the reverberation time to be controlled without changing the temporal structure of the reflections. The room was 6×3.5×2.6 m with microphone at [2.0, 1.0, 1.8] m and source at [2.866, 1.5, 1.8] m. The speed of sound was 348 m/s and the sample rate 16000 Hz. The clean speech signal was delayed to match the time of arrival of the direct sound and the epochs estimated using YAGA [11]. The LP filter $1/A_i(z)$ with order 18 was estimated for each larynx cycle of the delayed clean speech using the autocorrelation method over the interval $\kappa_i - {}^1\!/_2 L_i < n < \kappa_i + {}^3\!/_2 L_i$.

Each cycle of the delayed clean speech and the reverberant speech was filtered by its corresponding $A_i(z)$ to obtain the clean residual $e(n)$ and reverberant residual $\epsilon(n)$, respectively. Finally the larynx synchronous average was performed for each larynx cycle according to (13).

To evaluate the performance of the algorithm we define the A-weighted signal-to-distortion ratio as

$$\mathrm{SDRA}_y = 10 \log_{10} \frac{\sum_n \left[\mathcal{A}\left\{ d(n) \right\}^2 \right]}{\sum_n \left[\mathcal{A}\left\{ d(n) - y(n) \right\}^2 \right]} \qquad (14)$$

where $d(n)$ is the desired signal, $y(n)$ is the distorted signal and $\mathcal{A}\{\cdot\}$ is an A-weighting operation [12]. In all metrics we choose $d(n)$ to be $e(n)$. The effects of reverberation and its reduction by larynx synchronous averaging are given by choosing $y(n) = \epsilon(n)$ and $y(n) = \bar{\epsilon}(n)$, respectively. The improvement in SDRA is then $\mathrm{SDRA}_{\bar{\epsilon}} - \mathrm{SDRA}_{\epsilon}$.

The averaging process also distorts the underlying clean speech residual. This is quantified by applying the averaging directly to the clean residual to give $\bar{e}(n)$ and again measuring the A-weighted signal-to-distortion ratio.

4.2. Effect of T_{60}

For the same room geometry, increasing the reverberation time spreads the energy due to each larynx cycle over more of the subsequent frames. The extension of the noise-like reverberation tail has the effect of reducing the coherence between successive cycles, thus reducing the correlation between the observed frames. Fig. 2 shows the effect of varying the T_{60} of the reverberant residual and on the averaged residual with $c^-{=}c^+{=}3$ (i.e. $N{=}7$) and $\mathbf{w}{=}\mathbf{w}_R$. It is clear that larger reverberation times lead to a greater improvement in the SDRA.

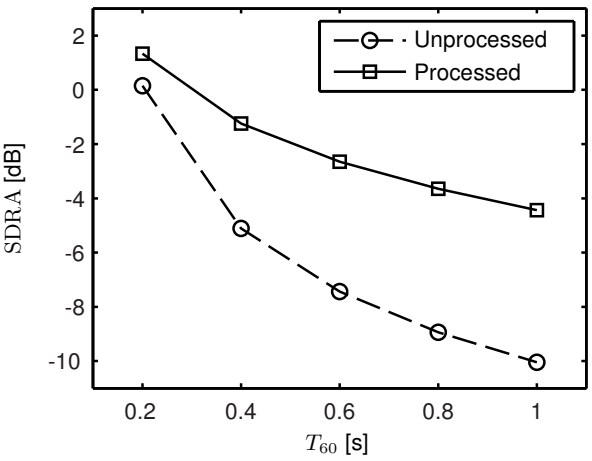

Fig. 2. Signal-to-distortion ratio as a function of reverberation time for unprocessed reverberant residual $\epsilon(n)$ and larynx synchronous averaged (processed) version $\bar{\epsilon}(n)$ with $c^-{=}c^+{=}3$ and rectangular window.

4.3. Effect of algorithm parameters

Using a fixed T_{60} of 0.6 s we consider the effect of the number of cycles included in the average for symmetrical (non-causal) averaging and causal averaging, as would be required to minimise latency in a real-time application. We also consider the effect of the Tukey window used in [4]. Fig. 3 shows that the attenuation of reverberation is substantial (>2.5 dB) even with only 3 cycles. Further improvements are achieved as N increases. The use of causal processing produces almost exactly the same amount of attenuation as symmetric processing. On the other hand, the Tukey window does limit the amount of dereverberation that can be achieved by between 0.5 and 1 dB depending on the T_{60}.

Fig. 4 shows the effect of the same parameter variations on the distortion of the clean speech residual, where for $N = 1$ $\mathrm{SDRA}_{\bar{e}} = \infty$ by definition. The most striking factor here is the improvement of approximately 2.5 dB that the Tukey window gives for speech quality. Informal listening suggests that this is a significant improvement which justifies the slight reduction in dereverberation.

5. DISCUSSION AND CONCLUSIONS

The signal model developed in Sec. 3 leads to (8), which expresses the reverberant residual as the summation of contributions from each larynx cycle of the clean speech residual. Larynx synchronous averaging across neighbouring cycles emphasises those contributions which are coherent between cycles. The contribution to the current observed frame from the current clean speech residual is always time-aligned and so sums coherently. The reverberation tail is stochastic in na-

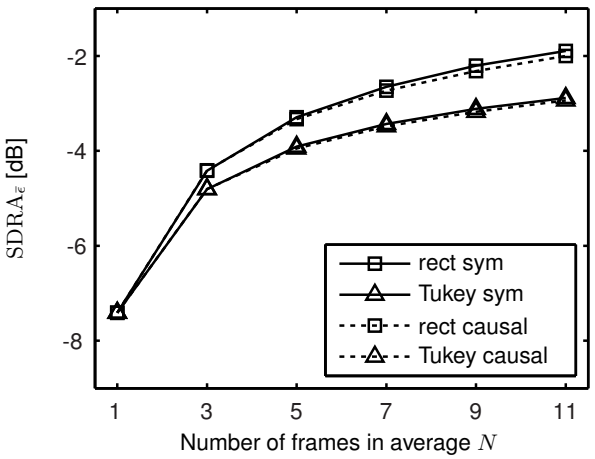

Fig. 3. Signal-to-distortion ratio of processed reverberant residual ($\mathrm{SDRA}_{\bar{\epsilon}}$) as a function of N for rectangular and Tukey windows and for causal and non-causal processing. SDRA_{ϵ} is represented as N=1.

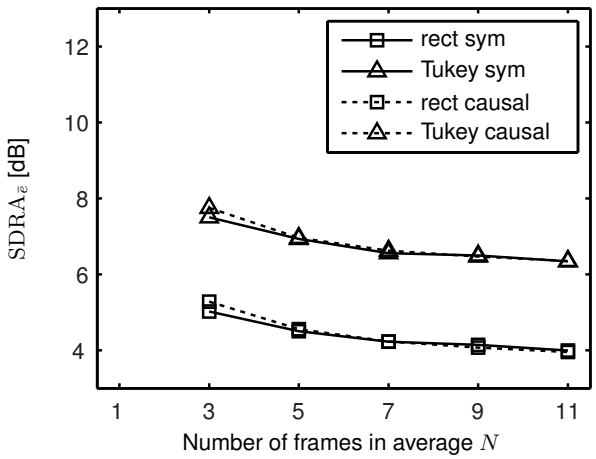

Fig. 4. Signal-to-distortion ratio of processed clean residual ($\mathrm{SDRA}_{\bar{\epsilon}}$) as a function of number of larynx cycles included in average for rectangular and Tukey windows and for causal and non-causal processing.

ture and so sums incoherently. Thus, intercycle averaging of voiced speech achieves dereverberation.

Simulation results have shown that for the same room geometry longer reverberation times lead to greater attenuation of the reverberation. The amount of dereverberation achieved increases with the number of cycles included in the average. Causal processing, as required for real time applications, does not compromise the dereverberation performance, when compared with non-causal processing with the same total number of cycles. Excluding the glottal closures from the averaging using a Tukey window slightly reduces the amount of dereverberation but achieves lower speech distortion.

REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[2] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[3] N. D. Gaubitch, M. R. P. Thomas, and P. A. Naylor, *Dereverberation using LPC-based approaches*, chapter 4, pp. 95–128, Springer, 2010.

[4] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Cardiff, UK, July 2007.

[5] M. R. P. Thomas, N. D. Gaubitch, J. Gudnason, and P. A. Naylor, "A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2007.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[7] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[8] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Multichannel DYPSA for estimation of glottal closure instants in reverberant speech," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Poznan, Poland, Sept. 2007.

[9] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transforms," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[11] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Detection of glottal opening and closing instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 82–91, 2012.

[12] IEC, "IEC 61672:2003: Electroacoustics – sound level meters," Tech. Rep., IEC, 2003.