ACOUSTIC BLUR KERNEL WITH SLIDING WINDOW FOR BLIND ESTIMATION OF REVERBERATION TIME

Felicia Lim, Patrick A. Naylor

Dept. of Electrical and Electronic Engineering
Imperial College London, UK
{felicia.lim06, p.naylor}@imperial.ac.uk

Mark R. P. Thomas, Ivan J. Tashev

Microsoft Research
Redmond, WA 98052, USA
{markth, ivantash}@microsoft.com

ABSTRACT

Reverberation time, or T_{60}, is a key parameter used for characterizing acoustic spaces. Blind T_{60} estimation is useful for many applications including speech intelligibility estimation, acoustic scene analysis and dereverberation. In our previous work, a single-channel blind T_{60} estimator was proposed employing spectral analysis in the modulation frequency domain. It was shown that the estimation accuracy is crucially affected by the window lengths used for transformation to the modulation domain. In this work, we propose the use of a sliding window length that is dynamically updated depending on the length of the detected decay region. Experimental results demonstrated that in the presence of noise, estimation accuracy was improved over our previous work for T_{60} up to 700 ms. When compared against two alternative algorithms from the literature, the proposed approach demonstrated higher accuracy for T_{60} between 500 ms and 1 s. Finally, the proposed approach was shown to be more computationally efficient compared to two of the three alternative algorithms.

Index Terms— blind reverberation time estimation

1. INTRODUCTION

Introduced by Sabine in the late 1890s [1], reverberation time is defined as the time taken for the energy of a steady-state sound field to decay by 60 dB after the excitation source signal has been switched off [1]. The reverberation time, also common referred to as T_{60}, is a function of the room geometry and the reflectivity of all surfaces within, but is independent of the source-receiver geometry. Knowledge of this acoustic parameter is interesting for many acoustic applications such as speech intelligibility estimation, robust automatic speech recognition, acoustic scene analysis and dereverberation. If a room's acoustic impulse response (AIR) is available, its T_{60} can be measured using Schroeder's backward integration method [2]. This method calculates the energy decay curve [3] of the AIR and applies a linear fit to the region of free decay, typically selected to be between -5 and -35 dB, depending on the noise floor.

In practical scenarios, measured AIRs are not always easily obtainable and therefore it is desirable to estimate T_{60} directly from the reverberant, and often noisy, signals captured at the microphone(s). In [4, 5], neural network approaches were developed using samples of the time domain reverberant signal and speech envelope power spectral densities respectively while in [6], gaps in the speech signal were identified to track the decay curve. Another time domain approach employs maximum likelihood (ML) estimation [7] and was improved upon in [8] to reduce computational complexity and increase robustness to background noise. In the frequency domain, the spectral decay distribution (SDD) method was proposed in [9] and improved in [10], where frequency-dependent decay rates of reverberant speech are estimated and the negative-side variance of its histogram is mapped to T_{60}. In the modulation frequency domain, the speech-to-reverberation modulation ratio (SRMR) method was proposed based on the smearing of reverberant energy in the modulation domain and its inverse was shown to be highly correlated with T_{60} [11]. More recently, the reverberation problem was investigated as an image blurring problem in [12] and the T_{60} was derived from the estimation of the blur kernel's parameters, also in the modulation domain.

A comparative evaluation of [8, 9, 11] was conducted in [13], where it can be seen that, even in the noise-free case, there is room for improvement in estimation accuracies, especially at higher T_{60}. In [12], it was shown that the proposed blur kernel estimation approach was able to improve accuracy for larger T_{60}. However, its cascade approach to estimating high and low T_{60} values separately results in large estimation errors in the cross-over region. Futhermore, the algorithm requires multiple iterations over the same speech segment, which is computationally expensive.

This work builds upon [12] and proposes the use of a sliding window length for detecting decay regions in the speech signal. This removes the need for a cascade approach previously adopted, which significantly reduces computational complexity. Experimental results demonstrated an improvement in estimation accuracy in the presence of noise for $T_{60} <\approx 700$ ms compared to [12].

The remainder of the paper is organized as follows. In Section 2, the reverberation problem is introduced as an image blurring problem and in Section 3, an overview of the blur kernel estimation algorithm from [12] is provided. The proposed modification is then presented in Section 4 and experimental results are given in Section 5. Finally, some conclusions are drawn in Section 6.

2. ACOUSTIC BLUR KERNEL

A reverberant signal is obtained as the linear convolution between clean speech $s[n]$ and an AIR $h[n]$, given as

$$x[n] = s[n] * h[n], \tag{1}$$

where $n \geq 0$ is the discrete time index and the sampling frequency is denoted f_s. The late reverberant tail of the AIR can be modelled as a non-stationary stochastic process [14]

$$h[n] = b[n]e^{-\alpha n}, \tag{2}$$

where $b[n]$ is a zero-mean stationary Gaussian noise and the decay rate is related to the T_{60} by

$$\alpha = 3\log(10)/T_{60}. \tag{3}$$

The short time Fourier transform (STFT) of $s[n]$ is given as

$$S[m,k] = \sum_{n=-\infty}^{\infty} s[n] w_{\text{a}}[n - pm] e^{-2\pi i k n / L}, \tag{4}$$

where m is the STFT frame index, k is the frequency bin, $w_{\text{a}}[n]$ is a causal window function of support L samples and p is the frame increment in samples. The STFT of reverberant speech, $X[m,k]$, can be obtained in a similar manner. This work is concerned only with the magnitude spectra of $S[m,k]$ and $X[m,k]$ and therefore the signals of interest are real. Their log magnitudes are plotted in Fig. 1, where it can be seen that the exponential reverberant tail has the effect of smearing energy into subsequent time frames. This

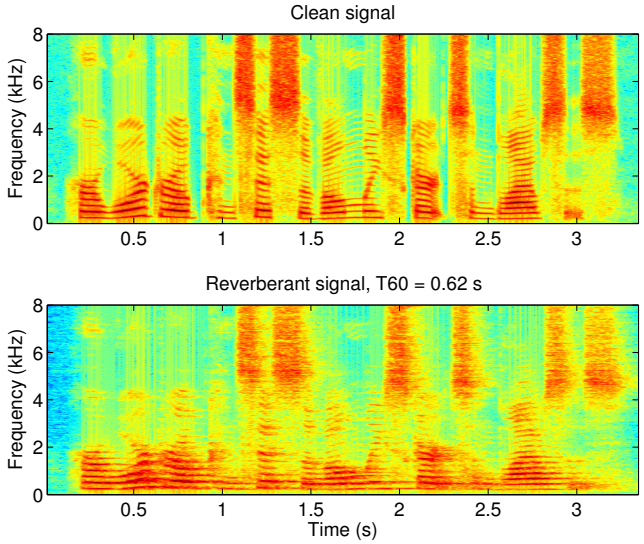

Figure 1: Spectrograms of clean and reverberant speech signals, where the T_{60} of the reverberant signal is 620 ms.

effect is analogous to motion blur in images, which may be modelled in a similar way to (1), where $s[n]$ would denote the original high resolution image and $h[n]$ is usually termed the blur kernel in 2D. In image processing, the topic of blur kernel estimation is well-studied for deblurring [15, 16, 17]. In acoustic signal processing, and particularly for blind T_{60} estimation, such an estimated blur kernel is interesting as an estimated T_{60} can be derived from it as $T_{60} = 3\log(10)/\alpha$.

3. BLUR KERNEL ESTIMATION

A common method of blur kernel estimation in image processing is through inspection of the blur kernel's Fourier transform to determine its direction and magnitude. In acoustic reverberation, the direction is known to always be along the time axis towards $n = +\infty$. Therefore, spectral analysis can be applied by simply taking the STFT of a reverberant signal in one direction across time.

Consider a simplified model of the AIR as $e^{-\alpha n}$ and the case when $s[n]$ is an impulse $\delta[n]$, i.e. $x[n] = \delta[n] * e^{-\alpha n}$. The STFT magnitude spectrum of $x[n]$ can be approximated as

$$|X[m,k]| \approx e^{-\alpha p m}, \tag{5}$$

since the discrete Fourier transform (DFT) of $\delta[n]$ is unity at each frequency.

In order to examine the behaviour of $|X[m,k]|$ as a function of time, a second STFT is applied with respect to m as

$$\tilde{X}[m',k',k] = \sum_{m=-\infty}^{\infty} |X[m,k]|\, w_{\text{mod}}[m - p'm'] e^{-2\pi i k' m / L'}, \tag{6}$$

where m' is the discrete time index in the modulation domain, k' is the modulation frequency, w_{mod} is the causal window function of support L' samples and p' is the frame increment in the acoustic frequency domain.

It is expected that the spectral analysis will yield the DFT of $e^{-\alpha p m}$ with respect to the STFT time index m, denoted $H_\alpha[k']$. This can be shown by considering the case where $L' \geq T_{60} f_{\text{s}}$ such that the signal decay in the region of interest for T_{60} estimation is captured within the first frame $m' = 0$. Therefore, (6) can be simplified to

$$\begin{aligned}\tilde{X}[0,k',k] &= \sum_{m=-\infty}^{\infty} |X[m,k]|\, w_{\text{mod}}[m] e^{-2\pi i k' m / L'} \\ &\simeq \sum_{m=0}^{L'-1} e^{-\alpha p m} e^{-2\pi i k' m / L'} = H_\alpha[k'],\end{aligned} \tag{7}$$

which is the DFT of $e^{-\alpha p m}$.

The blur kernel estimator [12] then estimates the decay rate by finding an α that results in a best fit of the magnitudes $|H_\alpha[k']|$ to $|\tilde{X}[m',k',k]|$ in the least-squares sense. For broadband T_{60} estimation, $|\tilde{X}[m',k',k]|$ is first averaged over all acoustic frequencies k to yield $|\tilde{X}[m',k']|$. Then, α is estimated as

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ \frac{1}{L'} \sum_{k'=0}^{L'-1} \left| H_\alpha[k'] - \tilde{X}[m',k'] \right|^2 \right\} \tag{8}$$

and used to derive the T_{60} estimate, denoted $\hat{T}_{60}$, for frame m'.

4. VARIABLE WINDOW LENGTH

For practical implementation, it was shown in [12] that the choice of window lengths L and L' is crucial to the accuracy of T_{60} estimation, where longer window lengths are desirable for high T_{60} while shorter window lengths are desirable for low T_{60}. A cascade approach was therefore adopted in [12] where two sets of longer window lengths were first used for high T_{60} estimation, followed by a set of shorter window lengths if the estimated T_{60} was smaller than a given expected range. This approach requires up to 3 iterations over the microphone signal, which is computionally expensive. Additionally, in real-world scenarios, both speech and AIRs contain spectral components that cause deviation of $|\tilde{X}[m',k']|$ from the ideal $|H_\alpha[k']|$. To mitigate this, [12] employed a signal pre-selection stage that attemps to determine if signal decay is present over the entire duration of a fixed-length frame m'. Misclassification can occur if signal decay only occurs in the first part of the frame and therefore a more flexible approach is desirable.

In this work, the use of sliding window lengths for both the decay detection stage and L' is proposed. Let L'_{max} denote the empirical maximum length allowed for L' such that the m'-th frame is constructed as

$$\mathbf{X}_{m',k} = \left[X[p'm',k] \;\ldots\; X[p'm' + L'_{\text{max}} - 1, k] \right]^T. \tag{9}$$

We wish to extract the region of $x[n]$ corresponding to $\mathbf{X}_{m',k}$ in order to determine if the signal is decaying within this frame. Firstly, consider $X[m,k]$, which is computed from

$$\mathbf{x}_m = [x[mp]\ x[mp+1]\ \ldots\ x[mp+L-1]]\,. \tag{10}$$

Then, the region of $x[n]$ corresponding to $\mathbf{X}_{m',k}$ can be found as

$$\mathbf{x}_{m'} = \big[[x[pp'm']\ x[pp'm'+1]\ \ldots \\ x[p(p'm'+L'_{\text{max}}-1)+L-1]\big]. \tag{11}$$

A dynamic decay detector is then applied to $\mathbf{x}_{m'}$ by firstly dividing the frame into Q subframes of length L_{q}. The subframed signals are denoted as $\mathbf{x}_{q,m'}$, for $q = 0,\ldots,Q-1$. The variance, maximum and minimum values were subsequently computed for each subframe. In a similar fashion to [8, 12], the consecutive subframes $\mathbf{x}_{q,m'}$ and $\mathbf{x}_{q+1,m'}$ are classified as decaying regions if the following are true

$$\text{var}\{\mathbf{x}_{q,m'}\} > \text{var}\{\mathbf{x}_{q+1,m'}\}, \tag{12a}$$

$$\max\{\mathbf{x}_{q,m'}\} > \max\{\mathbf{x}_{q+1,m'}\}, \tag{12b}$$

$$\min\{\mathbf{x}_{q,m'}\} < \min\{\mathbf{x}_{q+1,m'}\}. \tag{12c}$$

If a non-decaying subframe $\mathbf{x}_{q+1,m'}$ is detected, and $q \geq q_{\text{min}}$, where q_{min} is a pre-defined minimum number of subframes, then the signal up to the q-th frame is extracted as a decaying frame, where its length in the time domain is $N = (q+1)L_{\text{q}} - 1$.

We now want to find a value for L' that is sufficiently long to capture only the detected decaying region of $\mathbf{X}_{m',k}$. Since L' denotes the number of frames in the acoustic frequency domain, it can be found as the largest $L' \in \mathbb{Z}^+$ that satisfies

$$L'p + L - 1 < N, \tag{13}$$

and the corresponding decaying region in the acoustic frequency domain is obtained as

$$\check{\mathbf{X}}_{m',k} = \big[X[p'm',k]\ \ldots\ X[p'm'+L'-1,k]\big]^T. \tag{14}$$

Equation (6) can now be applied, followed by estimation of α using (8) and computation of its corresponding T_{60} for the m'-th frame. In this manner, a variable L' allows a more flexible decay detector to be implemented and eliminates the need for a cascade approach employing multiple combinations of different L and L'.

After all decaying frames have been processed, the T_{60} estimates are averaged across all frames, yielding the final estimate. A summary of the algorithm is provided in Algorithm 1.

Algorithm 1 Blur kernel with sliding window

```
Compute X[m,k].
for all frames m' do
    Compute X_{m',k} using L'_max as in (9).
    Find the corresponding time domain region, x_{m'}, as in (11).
    Divide x_{m'} into Q subframes.
    Detect possible sound decay within x_{m'} according to (12).
    if decaying region is found then
        Find the largest L' ∈ Z^+ that satisfies (13).
        Compute (6).
        Find α that minimizes (8).
        Compute and store the corresponding T̂_60.
    end if
end for
Average T̂_60 over all frames to find the final estimate.
```

5. EVALUATION

The proposed blur kernel with sliding window was evaluated against the blur kernel approach [12] and two additional algorithms from the literature: 1) ML [8], and 2) SRMRinv [11]. Reverberant signals were obtained by convolving 10 clean speech files from the TIMIT database with 21 measured AIRs from the AACHEN database, yielding 210 signals. Noise was added as white Gaussian noise (WGN) with signal-to-noise ratios (SNRs) $\in \{20, 10\}$ dB. The ground truth T_{60} values were measured from the AIRs using Schroeder's backward integral, where the free decay regions were fitted by hand. The parameters in the proposed algorithm were determined empirically as follows. The length L was fixed at 128 ms and L'_{max} was set to 350 ms to capture the longer decays for high T_{60}. The window function used for $w_{\text{a}}[n]$ and $w_{\text{mod}}[m]$ was the square-root of a periodic Hann window and the frame increments were arbitrarily chosen as $p = L/4$ and $p' = 1$. In the decay detection algorithm, $q_{\text{min}} = 4$ and $L_{\text{sf}} = 34$ ms were used. The estimation error was calculated as

$$\varepsilon = \hat{T}_{60} - T_{60}, \tag{15}$$

and the results are given in Fig. 2.

Algorithm	Averaged real time factors
ML	0.02
SRMRinv	0.32
Blur Kernel	6.66
Blur Kernel with Sliding Window	0.23

Table 1: Averaged real time factors.

When comparing the proposed approach against the blur kernel, it can be seen that the proposed approach demonstrated smaller variances in general. In the case of no noise (SNR $= \infty$ dB) , the proposed algorithm improves the estimation accuracy for T_{60} in the range of approximately 500 to 700 ms, but slightly reduced median estimation errors at lower T_{60} with differences of $\leq$ 100 ms. In the presence of noise, the proposed algorithm maintained similar or improved T_{60} estimation up to $\approx$ 700 ms, with the improvement becoming more significant as the SNR decreases. However, it demonstrated reduced accuracy for higher $T_{60} > \approx 800$ ms over all SNRs. These inaccuracies appear to arise from inaccurate decay detection and therefore this remains an area of open research. When comparing against ML and SRMRinv, it can be seen that the proposed approach achieved smaller estimation errors for $T_{60} > \approx 500$ over all SNRs but was not as robust for smaller T_{60}, especially at lower SNR.

Additionally, the computational complexity of all algorithms were evaluated by computing their real time factors, given in Table 1. The proposed method demonstrated a smaller real time factor compared to SRMRinv and blur kernel, while ML is the least computationally expensive algorithm.

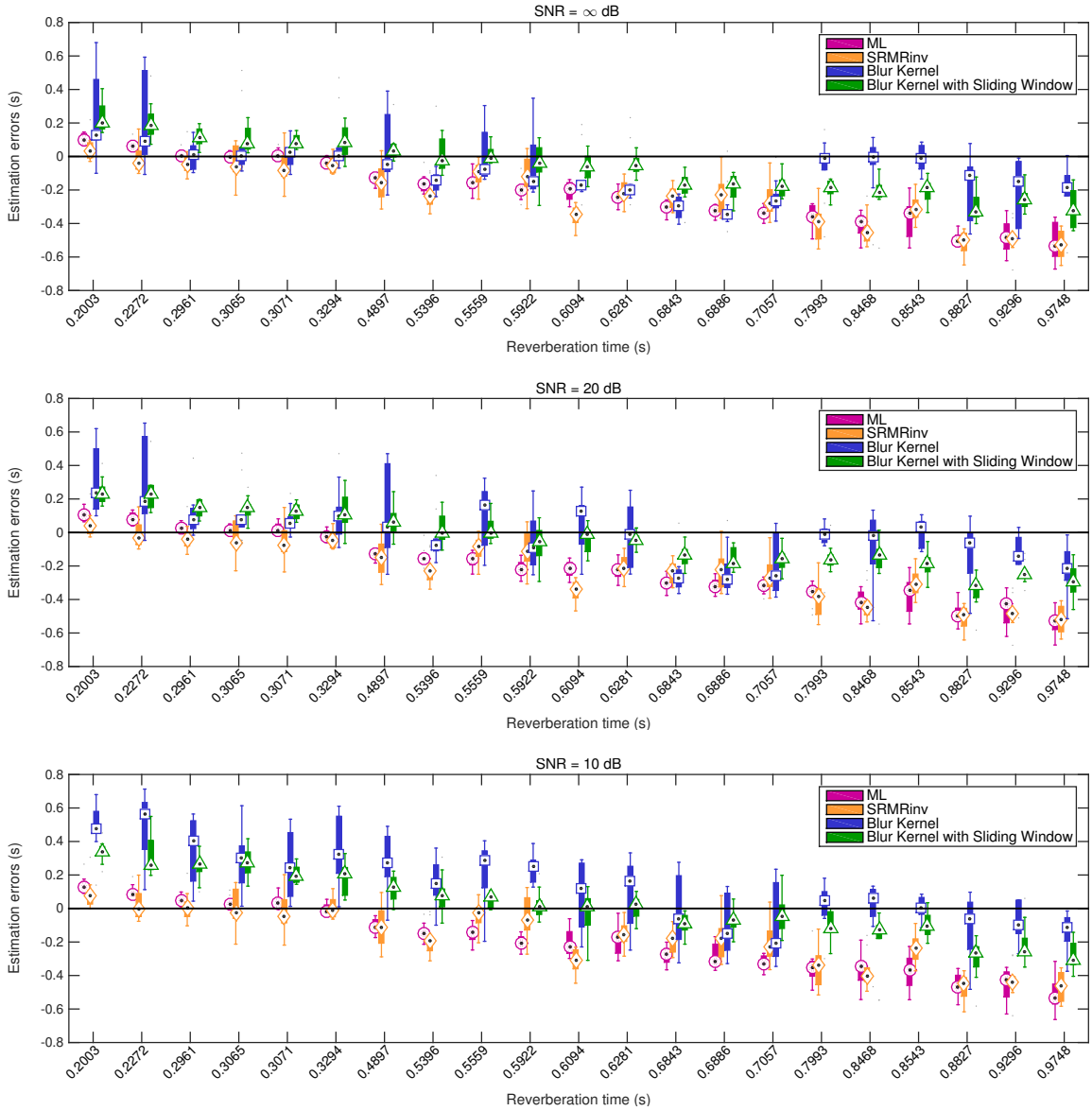

Figure 2: Estimation errors for the four algorithms compared, as a function of T_{60} and for different SNRs. The black dots denote the median, the thick vertical lines show the interquartile ranges and the thin vertical lines indicate the range up to 1.5 times the interquartile range.

6. CONCLUSIONS

This work extends the blur kernel algorithm [12] where the blind T_{60} estimation problem was approached from an image processing perspective and spectral analysis was employed in the modulation frequency domain. It was shown in [12] that careful selection of window lengths used for transforming the microphone signal into the modulation frequency domain was crucial for estimation accuracy. We now propose the use of a sliding window length that is dependent on the length of detected decay regions in the noisy and reverberant signals. Evaluation was carried out on noisy reverberant signals, and it is shown that the proposed algorithm was able to improve estimation accuracy for T_{60} up to $\approx$ 700 ms. However, it suffers from larger estimation errors at high T_{60} and it was noted that imperfect signal decay detection contributed to these errors. Additional evaluation against two other algorithms from the literature showed that the proposed approach achieved smaller estimation errors at high T_{60} but were not as robust to noise at lower T_{60}. Finally, it was demonstrated that the proposed algorithm significantly reduced the real time factor of the blur kernel algorithm [12] and was additionally computationally faster than SRMRinv [11].

7. REFERENCES

[1] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000.

[2] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

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

[4] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *Journal Audio Eng. Soc.*, vol. 49, no. 4, pp. 219–230, 2001.

[5] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, 2003.

[6] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[7] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[8] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel-Aviv, Israel, Aug. 2010.

[9] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, Apr. 2008.

[10] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[11] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[12] F. Lim, M. R. P. Thomas, and I. J. Tashev, "Blur kernel estimation approach to blind reverberation time estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015.

[13] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012.

[14] J. D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, France, 1988.

[15] M. M. Chang, A. M. Tekalp, and A. T. Erdem, "Blur identification using the bispectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 39, pp. 2323–2325, Oct. 1991.

[16] C. Mayntz, T. Aach, and D. Kunz, "Blur identification using a spectral inertial tensor and spectral zeros," in *Proc. Intl. Conf. Image Processing*, Oct. 1999, pp. 885– 889.

[17] B. Kang, J. Shin, and P. Park, "Piecewise linear motion blur identification using morphological filtering in frequency domain," in *ICROS-SICE International Joint Conference*, Aug. 2009, pp. 1928–1930.