BLUR KERNEL ESTIMATION APPROACH TO BLIND REVERBERATION TIME ESTIMATION

Felicia Lim[1*], *Mark R. P. Thomas*[2] *and Ivan J. Tashev*[2]

[1]Dept. of Electrical and Electronic Engineering, Imperial College London, UK
[2]Microsoft Research, Redmond, WA 98052, USA
email: felicia.lim06@imperial.ac.uk, {markth, ivantash}@microsoft.com

ABSTRACT

Reverberation time is an important parameter for characterizing acoustic environments. It is useful in many applications including acoustic scene analysis, robust automatic speech recognition and dereverberation. Given knowledge of the acoustic impulse response, reverberation time can be measured using Schroeder's backward integration method. Since it is not always practical to obtain impulse responses, blind estimation algorithms are sometimes desirable. In this work, the reverberation problem is viewed as an image blurring problem. The blur kernel is estimated through spectral analysis in the modulation domain and the T_{60} is subsequently estimated from the blur kernel's parameters. It is shown through experimental results that the proposed approach is able to improve robustness to higher T_{60}s especially with increasing levels of additive noise up to an signal-to-noise ratio (SNR) of 10 dB.

Index Terms— blind reverberation time estimation

1. INTRODUCTION

In enclosed spaces, sound propagation from a source to a distant receiver may follow multiple paths due to reflections off surfaces within the room, resulting in a persistence of sound that is known as reverberation. The level of reverberation can be quantified by the reverberation time, or T_{60}, 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]. T_{60} is a function of room geometry and reflectivity of surfaces within the room [2], and therefore can be used to characterize the acoustic space. It is an interesting parameter for many applications including speech intelligibility estimates, robust automatic speech recognition, acoustic scene analysis, dereverberation and more.

The most commonly used method for measuring T_{60}, given knowledge of a room's acoustic impulse response (AIR), is through the use of Schroeder's backward integration method [3]. This method calculates the energy decay curve (EDC) [4] 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.

Since it is not always practical to obtain measured AIRs, blind methods for T_{60} estimation are desirable. In [5, 6], neural network approaches were developed using samples of the time-domain reverberant signal and speech envelope power spectral densities respectively. Another approach attempts to identify gaps in the speech signal to track the decay curve [7]. In [8], a maximum likelihood (ML) approach was developed and improved upon in [9] to reduce computational complexity and increase robustness to moderate background noise. More recently, the spectral decay distribution (SDD) method was proposed in [10] using frequency-dependent decay rates of reverberant speech in the short time Fourier transform (STFT) domain, and in [11], the reverberant-to-speech modulation ratio (RSMR) method is proposed based on the smearing of reverberant energy in the modulation domain. An evaluation of the latter three methods was conducted in [12], 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}s.

In this paper, the acoustic reverberation problem is viewed as an image blurring problem and a blind T_{60} estimator is proposed based on estimation of the blur kernel's parameters in the modulation domain. The performance is evaluated against the improved ML algorithm, SDD and RSMR.

The remainder of the paper is organized as follows. In Section 2, the reverberation problem is introduced as an image blurring problem. The proposed method for blind T_{60} estimation through blur kernel estimation is given in Section 3, followed by details for practical implementation in Section 4. Evaluation through experimental studies are given in Section 5 and conclusions are drawn in Section 6.

2. ACOUSTIC BLUR KERNEL

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

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

where $n \geq 0$ is the discrete time with an incremental step $1/f_s$ and f_s is the sampling frequency. The late reverberant tail of the AIR can be modelled as a non-stationary stochastic process [13]

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

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 STFT of $s[n]$ is given as

$$S[m,k] = \sum_{n=-\infty}^{\infty} s[n] w_a[n - n_a(m)] e^{-2\pi i k n / N_a}, \tag{4}$$

where $m \in \mathbb{Z}$ is the discrete time index in the STFT domain, k is the frequency bin, $w_a[n]$ is a window function of support L_a samples, the time at the mid-point of block m is $n_a(m) = mL_a(1 - 1/r)/f_s + (L_a - 1)/(2f_s)$ with r as the overlap factor, and N_a is the

[*] This work was carried out while at Microsoft Research.

number of discrete Fourier transform (DFT) points. 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 damping due to reverberation has the effect of smearing energy into subsequent time frames. This 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, or point spread function, in 2D. In image processing, the topic of blur kernel estimation is well-studied for deblurring [14, 15, 16]. 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 $T_{60} = 3\log(10)/\alpha$. In this work, one method of blur kernel estimation is explored for blind T_{60} estimation.

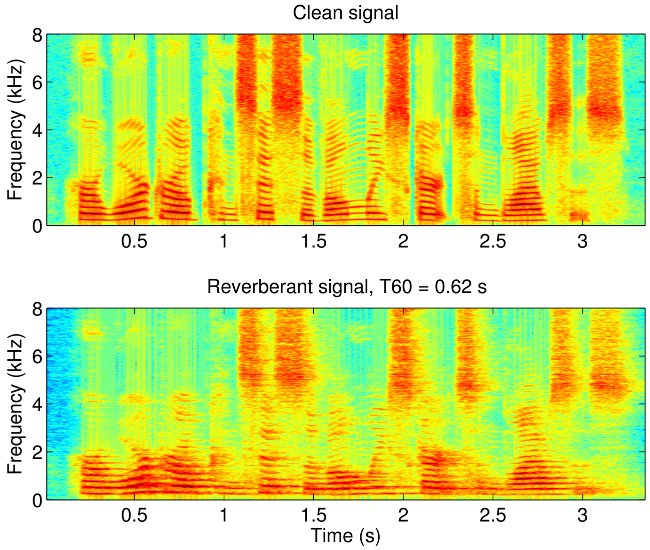

Fig. 1: Spectrograms of clean and reverberant speech signals, where the T_{60} of the reverberant signal is 620 ms.

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 spectra of $x[n]$ can be approximated as

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

since the DFT of $\delta[n]$ is unity at each frequency.

To conduct spectral analysis, a second STFT is applied to $|X[m,k]|$ in the direction of reverberation, i.e. across time frames and along each frequency bin, effectively transforming $x[n]$ into the modulation domain as

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

where $m' \in \mathbb{Z}$ is the discrete time index in the modulation domain, k' is the modulation frequency, w_{mod} is the window function of support L_{mod} samples and N_{mod} is the number of DFT points.

It is expected that the spectral analysis will yield the DFT of $e^{-\alpha n_a(m)}$ with respect to the STFT time index m, denoted $H[k']$. This can be shown by considering the case where $L_{\text{mod}} \geq T_{60} f_s$ 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

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

which is the DFT of $e^{-\alpha n_a(m)}$. The decay rate α can then be estimated by finding an α that results in the best fit of the magnitudes $|H[k']|$ to $|\tilde{X}[m',k',k]|$ in the mean squared error (MSE) sense as follows. As this work is concerned with broadband T_{60} estimation, $|\tilde{X}[m',k',k]|$ is first averaged over all acoustic frequencies k to yield $|\tilde{X}[m',k']|$. Then, the α that minimizes

$$e[m'] = \frac{1}{N_{\text{mod}}} \sum_{k'=0}^{N_{\text{mod}}-1} \left| H[k'] - \tilde{X}[m',k'] \right|^2 \tag{8}$$

is used to derive the T_{60} estimate for block m'.

4. PRACTICAL IMPLEMENTATION

In order to ensure successful practical T_{60} estimation by fitting $|H[k']|$ to $|\tilde{X}[m',k']|$, several issues must be addressed.

Firstly, realistic AIRs and source signals contain significantly more spectral components that cause deviation from the ideal $H[k']$. To mitigate this, signal pre-selection is performed as [9] such that only time frames containing possible signal decay are used for T_{60} estimation. Further details are given in Section 4.1. Additionally, a histogram is constructed from the T_{60}s estimated in all selected time frames and the modal histogram bin is selected as the final T_{60} estimate, $\hat{T}_{60}$.

Secondly, the choice of the two window lengths, L_a and L_{mod}, crucially affects the accuracy of T_{60} estimations. Selection of these window lengths is discussed in Section 4.2.

The proposed algorithm is finally summarized in Section 4.3.

4.1. Signal pre-selection

It is desirable to find frames where the contribution of speech is impulse-like, for example a hard plosive before a pause. To find such suitable frames, each frame of reverberant signal is passed through a sound decay detection stage to identify suitability for use in the T_{60} estimation stage. The criteria used for decay detection is similar to that proposed in [9] and is summarized as follows. Each frame $\tilde{X}[m',k']$ is divided into Q sub-frames, where the q-th subframe is denoted as $\tilde{X}[m',k',q]$ for $q = \{1,\ldots,Q\}$. The variance, maximum and minimum values in $x[n]$ corresponding to the time frame of the q-th subframe, denoted $x[n,q]$ is then compared with the same in the $(q+1)$-th subframe. The frame $\tilde{X}[m',k']$ is marked as containing possible sound decay and valid for use in T_{60} estimation if the following holds for all q:

$$\text{var}\{x[n,q]\} > k_v \cdot \text{var}\{x[n,q+1]\}, \tag{9a}$$

$$\max\{x[n,q]\} > k_{\text{mod}} \cdot \max\{x[n,q+1]\}, \tag{9b}$$

$$\min\{x[n,q]\} < k_{\text{mod}} \cdot \min\{x_[n,q+1]\}, \tag{9c}$$

where k_v and k_{mod} denote constant weighting factors.

4.2. Window lengths selection

Two important parameters affecting the accuracy of the proposed method are the two window lengths for transforming the time domain signal into the acoustic frequency domain (L_a) and for subsequent transformation into the modulation domain (L_{mod}). In the remainder of this paper, a combination of the two windows is denoted as $L = \{L_a, L_{\text{mod}}\}$. For higher T_{60}s, an L consisting of longer L_a and L_{mod} provides better estimates as more of the reverberant tail is captured within the longer time frames. For smaller T_{60}s, a combination of shorter window lengths is desirable since this limits the capture of noise floor remaining after the signal decay. It was found empirically that for T_{60}s in the approximate range of 400 to 900 ms, a combination of $L_a = \{0.016, 0.032, 0.064, 0.128\}f_s$ and $L_{\text{mod}} = \{0.25, 0.35\}f_s$ were more suitable for estimating T_{60}s to within ± 0.1 s. For T_{60}s in the approximate range of 100 to 600 ms, a combination of $L_a = \{0.008, 0.016\}f_s$ and $L_{\text{mod}} = \{0.15, 0.25\}f_s$ were found to be more suitable.

In order to handle the different window lengths required for good estimation of different T_{60}s, a 'cascade' approach is adopted where up to two iterations of the algorithm are run; the first iteration uses combinations of longer L to estimate the higher T_{60}s. If the final $\hat{T}_{60}$ is smaller than the lower threshold for the window length combinations used, the second iteration of the algorithm is run with shorter L. Details are given in Section 4.3.

4.3. Proposed algorithm summary

In this work, several parameters were chosen empirically, as follows. The window length combinations used for estimating higher T_{60}s are $L = \{0.064, 0.25\}f_s$ and $L = \{0.064, 0.35\}f_s$, while the combination used for estimating lower T_{60} values is $L = \{0.016, 0.15\}f_s$. The threshold value between these two T_{60} segements, used for switching to the second iteration in the 'cascade' approach, was set at 650 ms. For both $w_a[n]$ and $w_{\text{mod}}[m]$, the window function used was the square-root of a periodic Hann window and an overlap factor of $r = 2$ was used. In the decay detection algorithm, the following values were chosen: $N = 3$ when $L_m = 0.15f_s$, $N = 4$ when $L_{\text{mod}} = \{0.25, 0.35\}f_s$, $k_v = 0.9$ and $k_{\text{mod}} = 0.85$.

A summary of the proposed blur kernel algorithm is provided in Algorithm 1.

Algorithm 1 Proposed algorithm

```
$L_1 = \{0.064, 0.25\}f_s$, $L_2 = \{0.064, 0.35\}f_s$, $J = 2$
completed $=$ false, iteration2 $=$ false

while completed is false do
    for j = 1:J do
        Compute $\tilde{X}[m',k']$ using $L_j$
        for all frames in $\tilde{X}[m',k']$ do
            Detect possible sound decay.
            if current frame contains decay then
                Find $\alpha$ that minimizes (8).
                Compute and store corresponding $\hat{T}_{60}$.
            end if
        end for
    end for

    Group all $\hat{T}_{60}$s into bins of $0.1 : 0.02 : 1$ s.
    Find the final $\hat{T}_{60}$ as the modal bin.

    if final $\hat{T}_{60}$ $<=$ 650 ms and iteration2 is false then
        Clear all stored $\hat{T}_{60}$s.
        $L_1 = \{0.016, 0.15\}f_s$, $J = 1$
        iteration2 $=$ true.
    else
        completed $=$ true.
    end if
end while
```

5. EVALUATION

A total of 16 clean speech signals were taken from the TIMIT database, with different speakers and content, and 16 measured AIRs taken from the AACHEN database. The ground truth T_{60}s were measured using Schroeder's backward integral, where the region of free decay was fitted manually. Reverberant signals were obtained by convolving the clean speech signals and AIRs, giving a total of 256 reverberant signals. Noise was added as white Gaussian noise (WGN) with SNRs of $\{\infty, 30, 20, 10, 0\}$ dB. The resulting noisy and reverberant signals were not concatenated to avoid introducing multiple pauses in each speech sample. Evaluation was carried out by computing the estimation errors as $E = \hat{T}_{60} - T_{60}$.

The performance of the proposed algorithm was compared against the following three state-of-the-art T_{60} estimators: 1) ML [9], 2) RSMR [17], and 3) SDD [10]. Boxplots of the estimation errors computed for each algorithm and SNR considered are given in Fig. 2. It can be seen that in the noise-free case with SNR $= \infty$ dB, SDD and the proposed blur kernel algorithm demonstrate approximately equal, or increased robustness over ML and RSMR across the entire T_{60} range considered. In the region around 650 ms, the variance of estimation errors for the blur kernel increases significantly, due to the transition between the longer L and shorter L. This indicates that further investigation into the transition state would be useful to improve accuracy of the blur kernel estimator. In the presence of noise up to SNR $=$ 10 dB, the accuracy of blur kernel's estimation decreases at lower T_{60}s with increasing variance of estimation errors. However, in highly reverberant environments with $T_{60} \approx> 600$ ms, the blur kernel approach achieves improved estimation accuracy out of all algorithms considered. At SNR $=$ 0 dB, the blur kernel approach is no longer robust at any T_{60} and instead, appears to be consistently estimating the T_{60} as ≈ 0.9 s. The ML algorithm is similarly non-robust, as it consistently estimates the T_60 as ≈ 0.4 s (as it has done for all experimental scenarios considered). However, RSMR exhibits a sudden increase in estimation accuracy across all T_{60}s compared to higher SNRs.

6. CONCLUSIONS

Existing blind T_{60} estimators have been shown to deviate from the true T_{60} values especially in highly reverberant and/or noisy envi-

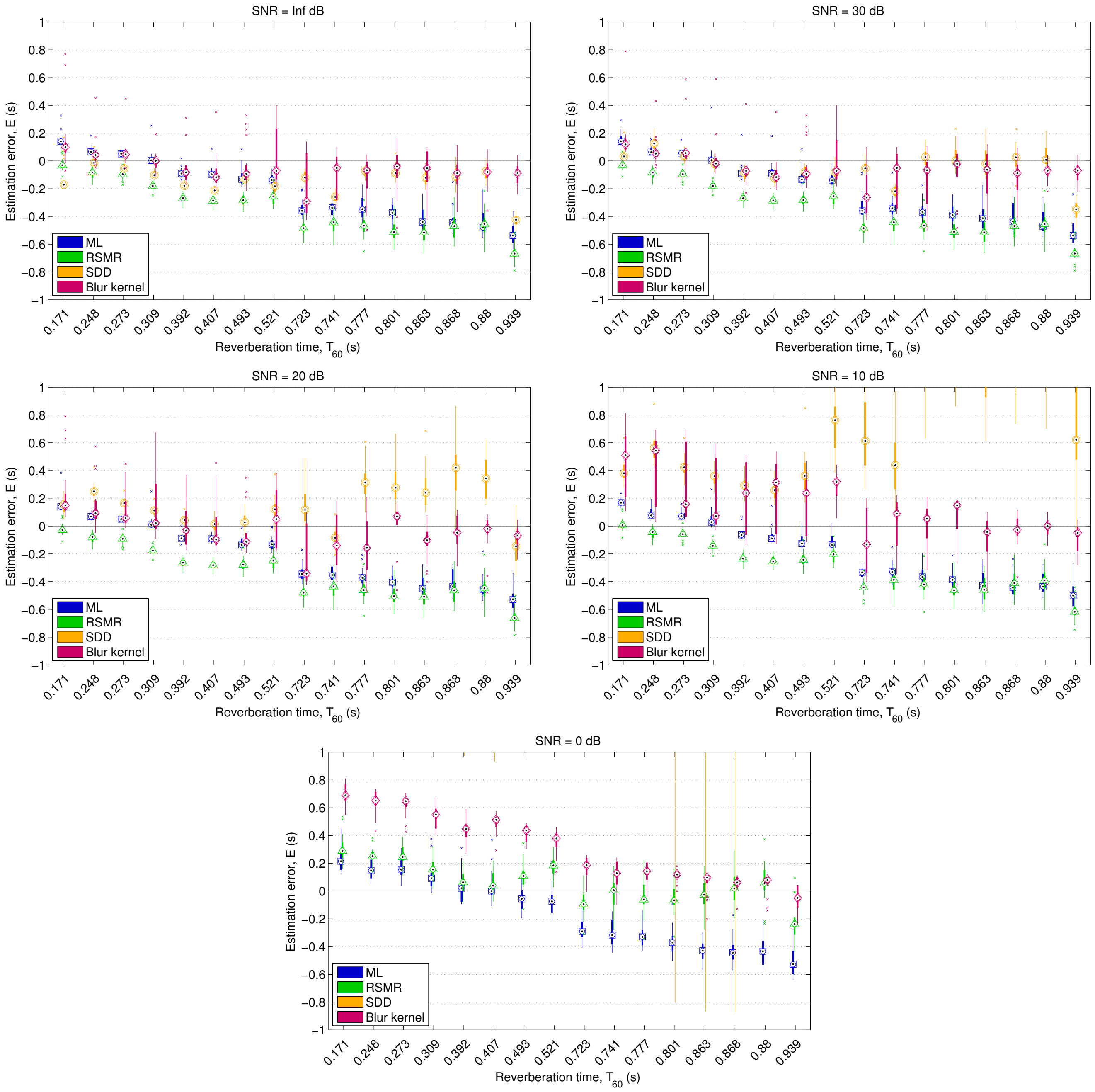

Fig. 2: Errors in reverberation time estimation from noisy reverberant speech signals. Each group of boxplots show the distribution of estimation errors for the four algorithms considered, for each T_{60}. The thick vertical lines show the interquartile ranges, the black dots denote the median and the thin vertical lines indicate the range up to 1.5 times the interquartile range.

ronments. In this work, the blur kernel approach is proposed based on spectral analysis of the captured microphone signal in the modulation frequency domain. Several algorithm parameters were chosen empirically here, and further investigation may yield better tuned parameters. Evaluation was carried out using short segments of clean speech from the TIMIT database and real AIRs from the AACHEN database in the presence of WGN of varying SNRs. It was shown that in the noise-free case, the proposed approach demonstrated improved robustness to higher T_{60}s while maintaining similar levels of accuracy compared to alternative algorithms considered. In increasingly noisy environments up to SNR $= 10$ dB, while the proposed approach exhibits reduced accuracy at lower T_{60}s, it was able to improve robustness at higher T_{60}s.

7. REFERENCES

[1] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, fourth edition, 2000.

[2] C. F. Eyring, “Reverberation time in ’dead’ rooms,” *J. Acoust. Soc. Am.*, vol. 1, no. 2A, pp. 168, 1930.

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

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

[5] 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.

[6] F. F. Li and T. J. Cox, “Speech transmission index from running speech: A neural network approach,” *J. Acoust. Soc. Am.*, vol. 113, pp. 1999–2008, 2003.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] 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.

[13] J. D. Polack, *La transmission de l’énergie sonore dans les salles*, Ph.D. thesis, Université du Maine, Le Mans, France, 1988.

[14] 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.

[15] 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.

[16] 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.

[17] T. H. Falk and W.-Y. Chan, “Temporal dynamics for blind measurement of room acoustical parameters,” *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, Apr. 2010.