14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

PREDICTIVE DECONVOLUTION AND KURTOSIS MAXIMIZATION FOR SPEECH DEREVERBERATION

David T. Fee, C.F.N Cowan, Stefan Bilbao and Izzet Ozcelik

The Sonic Arts Research Centre
The Queen's University of Belfast

ABSTRACT

A predictive deconvolution, based on the linear predictive (LP) residual of speech, is used to extract an estimate of the inverse of the minimum phase component of a room impulse response. This inverse is applied as a prefiltering stage to a kurtosis maximizing (KM) adaptive filter to equalise the remaining non-minimum phase component. It was found that this improved the stability and performance of the KM filter for male speech but it was found that when the first stage LP order was increased the performance improved for both male and female speech.

1. INTRODUCTION

Linear Prediction (LP) has been used extensively for blind deconvolution in the field of geophysics and speech coding. In geophysics *predictive deconvolution* (PD) is used to obtain an estimate of a layered earth model from seismic readings [1] while in speech coding it is used to model a time-varying vocal tract function [2]. However, its use in the field of acoustic dereverberation has been minimal due to the mixed-phase nature of rooms.

Recent techniques for speech dereverberation, in which an attempt is made to recover the anechoic speech signal, have used higher order adaptive filtering. Gillespie et al. [3] outlined a kurtosis maximization adaptive filtering algorithm for the dereverberation of speech using a Modulated Complex Lapped Transform (MCLT) subband filter. In this the subband filter coefficients were adapted to maximise the kurtosis of the LP residual. A significant reduction for perceived reverberation was reported. However the authors found that use of the kurtosis maximization required a large amount of data and often resulted in unstable adaptation that was highly dependant on the room impulse response.

Most room transfer functions are often non-minimum phase due to the late energy in the room impulse response (RIR) [4, 5] but can be modelled as the convolution of an equivalent minimum phase component with an all-pass component. In this paper a technique is outlined that exploits the minimum phase, all-pass model of room reverberation. A primary stage is proposed that extracts a minimum phase estimation of the room impulse response using high order predictive deconvolution. The resulting coefficients approximate the inverse of an all-pole room model [6] that is applied to speech as a prefilter before use with the kurtosis maximization (KM) adaptive filtering algorithm stage. It has been found that by decomposing the deconvolution in this manner the overall performance is improved as well as the stability of the kurtosis maximization stage.

In this paper predictive deconvolution for speech is described in Section 2 then the kurtosis maximization adaptive filtering algorithm is described in Section 3. The results of the KM filter applied to speech are discussed followed by a combined approach using PD as a preprocess. Finally the effect of LP order on the algorithm is described in Section 4.

2. PREDICTIVE DECONVOLUTION FOR SPEECH

A speech signal recorded in an enclosed space $x(n)$ can be modelled as the convolution of a speech signal $s(n)$ with a finite room impulse response $h_r(n)$ of length N such that

$$x(n) = \sum_{k=0}^{N-1} h_r(k)s(n-k) \tag{1}$$

which can be represented in the z-domain as

$$X(z) = H_r(z)S(z)$$

For the purposes of this paper it is assumed that $H_r(n)$ is time invariant and that, for the applications considered, noise is below percievable levels and therefore ignored. It is well understood that a speech signal can be modelled as an excitation signal $u(n)$ convolved with a time-varying all-pole filter whose short-time transfer function $H_s(z)$ is given as [2]

$$H_s(z) = \frac{G}{1 - \sum_{k=1}^{P_{speech}} \alpha_k z^{-k}}$$

where G is a gain parameter and $\{\alpha_k\}$ a vector of all-pole filter coefficients of length P_{speech} where P_{speech} is typically between 10 and 20 for speech coding. Speech is assumed to be a stationary signal over a short time period, typically 20ms, and the model can thus be applied on a block-by-block basis. The output signal in each block can be modelled in the z-domain as

$$\begin{aligned} X(z) &= H_r(z)S(z) \\ &= H_r(z)H_s(z)U(z) \end{aligned}$$

where $H_r(z)$ is the room transfer function and $U(z)$ is the speech excitation signal.

It can be assumed that the period corresponding to the first 50ms of the RIR is perceived as part of the direct speech signal [7, 8]. Therefore, by performing a short block-based linear prediction (*short LP* block in Figure 1) the short-time transfer function $H_s(z)$ can be removed without affecting the component of $H_r(z)$ perceived as reverberation. The resulting LP residual signal $\hat{x}(n)$ will approximate the excitation signal $u(n)$ convolved with the room impulse response $h_r(n)$ (the hat notation is used to indicate the short-time LP residual).

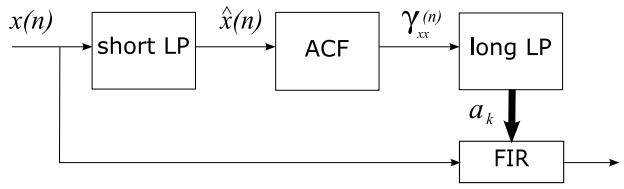

Figure 1: Predictive deconvolution for reverberant speech.

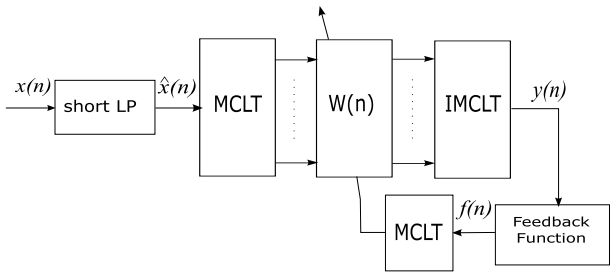

Figure 2: Kurtosis maximization of speech LP residual.

The room transfer function $H_r(z)$ is often non-minimum phase [8, 5] but can be modelled as the product of an equivalent minimum phase component $H_r^{min}(z)$ and all-pass component $H_r^{ap}(z)$ [5]:

$$H_r(z) = H_r^{min}(z)H_r^{ap}(z)$$

By assuming that the speech excitation signal $u(n)$ approximates a white noise process and the reverberated LP residual $\hat{x}(n)$ is a stationary stochastic process, performing a second linear prediction on a longer sequence of $\hat{x}(n)$ will give the all-zero system $A(z)$ whose coefficients approximate the reciprocal of an all-pole model of $H_r^{min}(z)$ with parameters $a_k, \;\; k = 1, \cdots, P_{long}$ [6] such that

$$H_r^{min}(z) \approx \frac{G}{1+\sum_{k=1}^{P_{long}} a_k z^{-k}} = \frac{G}{A(z)}$$

The resulting system $A(z)$ is then applied as an FIR filter to the reverberant speech signal $x(n)$ thereby deconvolving the minimum phase component of the RIR.

Although the RIR sequence is often very long (e.g. 4000 samples for a reverberation time of 0.5s and 8kHz sampling rate) fewer coefficients are required for an all-pole room model and although $H_r(z)$ will contain poles and mixed phase zeros, with a high enough prediction order the all-pole model will give a good approximation of H_r^{min} [6].

The filter coefficients of $A(z)$ are found by calculating the autocorrelation sequence $\gamma_{xx}(n)$ of the short-time LP residual $\hat{x}(n)$ and solving for the coefficients using the Levinson-Durbin recursion algorithm. This allows for the efficient solution of very large order predictions. This process is indicated by *long LP* block in Figure 1. $A(z)$, the estimated inverse of $H_r^{min}(z)$ is then applied as an FIR filter. The estimation of $A(z)$ can be improved by increasing the length of the autocorrelation sequence. For this reason the calculation of $A(z)$ is done *off-line*.

The technique used for extracting an estimate of the inverse of the minimum phase component of the RIR outlined above can be illustrated by the block diagram shown in Figure 1 where the input $x(n)$ is the recorded speech signal given by (1), $\hat{x}(n)$ is the LP speech residual, *ACF* represents autocorrelation calculation and $\gamma_{xx}(n)$ the autocorrelation sequence of $\hat{x}(n)$.

3. SUBBAND KURTOSIS MAXIMIZATION ADAPTIVE FILTERING

Since the predictive deconvolution stage outlined above is restricted to suppressing only the minimum phase component of the RIR a second stage capable of dealing with the remaining mixed-phase system is needed. A kurtosis maximizing adaptive filter fulfils this requirement allowing for an acausal filter to equalise the maximum phase room component.

The update equation for a kurtosis maximizing adaptive filter with coefficients $\mathbf{w}(n)$ at time n is given as

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu f(n)\hat{\mathbf{x}}(n)$$

where

$$f(n) = \frac{4\left[E\left\{\hat{y}^2(n)\right\}\hat{y}^2(n) - E\left\{\hat{y}^4(n)\right\}\right]\hat{y}(n)}{E^3\left\{\hat{y}(n)\right\}} \qquad (2)$$

in which μ is the adaptive step-size, $f(n)$ is a feedback function given by (2), $\hat{\mathbf{x}}(n)$ is the LP residual input vector and the output $\hat{y}(n) = \mathbf{w}^T(n)\hat{\mathbf{x}}(n)$. The terms $E\left\{\hat{y}^2(n)\right\}$ and $E\left\{\hat{y}^4(n)\right\}$ are estimated recursively as

$$E\left\{\hat{y}^2(n)\right\} = \beta E\left\{\hat{y}^2(n-1)\right\} + (1-\beta)\hat{y}^2(n)$$
$$E\left\{\hat{y}^4(n)\right\} = \beta E\left\{\hat{y}^4(n-1)\right\} + (1-\beta)\hat{y}^4(n)$$

with β controlling the smoothness of the estimates [3].

The subband structure is implemented using the Modulated Complex Lapped Transform (MCLT). The implementation updates are identical except that the update function $f(n)$ is calculated from an output block in the MCLT domain. Figure 2 shows a block diagram representation of the MCLT subband KM filter with the block-based short LP residual extraction as the first stage (*short LP*) and *IMCLT* representing the inverse MCLT.

3.1 KM applied to single channel speech

An experiment was carried out to measure the performance of this algorithm using male and female speech samples and a simulated room impulse response. Speech samples were taken from audio book CDs that were judged to be practically anechoic. These had a sampling rate of 44.1kHz and were decimated to 11025Hz. A non-minimum phase room impulse response was synthesised using the image method [9] to model a large sized room with dimensions 16m by 10m and height 9m in the x, y and z directions and a reverberation time of approximately 0.5s. Source and receiver coordinates in metres were $[2, 5, 2]$ and $[6, 5, 1.7]$ respectively. Figure 3 shows the resulting room impulse response. The number of MCLT subbands M is 1024 with a single tap in each subband adaptive filter.

The short-time LP residual was calculated with a pole order P_{speech} of 20 and an overlapping window of 256 samples and hop size of 128. The filter performance was measured

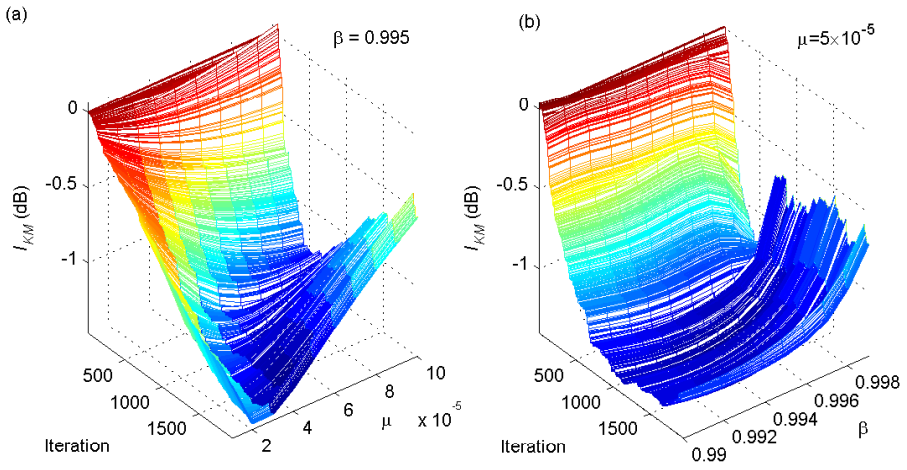

Figure 4: Dereverberation performance of the KM adaptive filter for male speech with increased iteration for values of μ and β: (a) β is constant at 0.995 and μ varied, (b) constant step-size $\mu = 3\times10^{-5}$ and β varied.

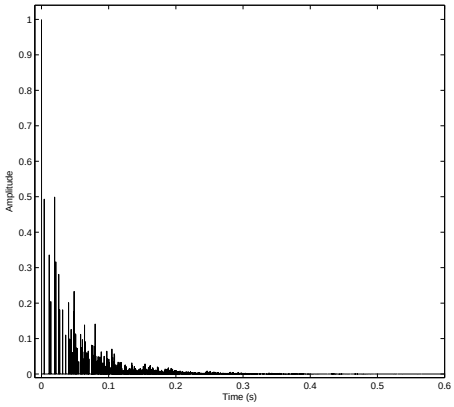

Figure 3: Image method synthesised room impulse response.

by calculating the reduction in reverberant energy when the resulting filter impulse response $w(n)$ was applied to the room impulse response. The reverberant error energy in the equalised system can be calculated as

$$R_c = \frac{\sum_n |c(n)|^2 - \max|c(n)|^2}{\max|c(n)|^2} \qquad (3)$$

where $c(n)$ is the equalised system vector given by

$$c(n) = h_r(n) * w(n)$$

($*$ denotes convolution). The maximum value of $\mathbf{c}$ is the direct signal component. The reverberant error energy R_h in the original room impulse response vector $\mathbf{h}_r$ is calculated in a similar manner where the maximum value of $h(n)$ is the direct signal component. The overall reduction in reverberant error energy in decibels or dereverberation performance is then

$$I(dB) = 10\log_{10}\{R_c\} - 10\log_{10}\{R_h\} \qquad (4)$$

such that $I = 0$ indicates no net improvement in reverberant energy.

By plotting I_{KM} against iteration the behaviour of the KM adaptive filter could be monitored[1]. It was found that the stability of the KM adaptive filter was greatly dependent on the values of μ and β. While certain combinations gave an improved dereverberation performance they often exhibited divergence. By varying both $\mu \in \{1,2,\cdots,10\}\times10^{-5}$ and $\beta \in \{0.99, 0.991, \cdots, 0.999\}$ optimal values could be estimated. For male speech the optimal values for this specific RIR were estimated as $\mu = 3\times10^{-5}$ and $\beta = 0.995$. Figure 4 shows two three-dimensional plots illustrating dereverberation performance with varied step-size μ and moment smoothing constant β. The instability of the KM filter can be seen on the right-hand-side of each plot after an optimal performance has been reached.

3.2 Predictive deconvolution as a KM prefiltering stage

The predictive deconvolution stage was tested as a prefiltering stage to the kurtosis maximization filter (Figure 5). The same speech samples were processed using the steps outlined in Section 2 (marked *Stage 1* in Figure 5) and passed through the resulting FIR filter. The output $\overline{x}(n)$ was used as the input to the KM adaptive filter. Both short LP residual extraction stages (*short LP 1* and *short LP 2* in Figure 5) had the same parameters as in the previous experiment.

The autocorrelation sequence was calculated from 60s of LP residual speech (661500 samples) and the predictive filter length P_{long} chosen as 2500, again $P_{speech} = 20$. The resulting reverberation reduction before the KM stage was found to be approximately $I_{PD} = -0.7$dB for male speech. The KM processing stage was applied using the same parameters. The complete equalised system was calculated after each iteration as

$$c(n) = h_r(n) * (a_{avg}(n) * w(n))$$

and the performance calculated as per Equations (3-4). The impulse response $w(n)$ of the KM filter was used in a second

[1]The subscript *KM* indicates the performance is measured for the *kurtosis maximization* stage only. *PD* represents the *predictive deconvolution* stage and *PDKM* indicates the combined approach.

FIR filter after convergence and dereverberated speech $y(n)$ extracted by filtering the first stage output $\bar{x}(n)$.

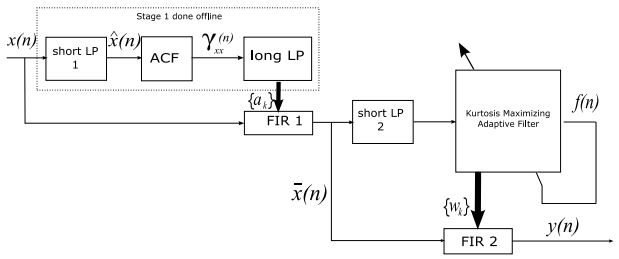

Figure 5: Block diagram showing predictive deconvolution stage as preprocess to the KM adaptive filter.

The resulting adaptive filter performance is plotted on Figure 6 alongside the *KM only* results (dotted and dot-dashed lines respectively). From these it can be seen that the combined filter converges faster and gives better performance for male speech with a reverberation reduction of almost $I_{PDKM} = -4$dB.

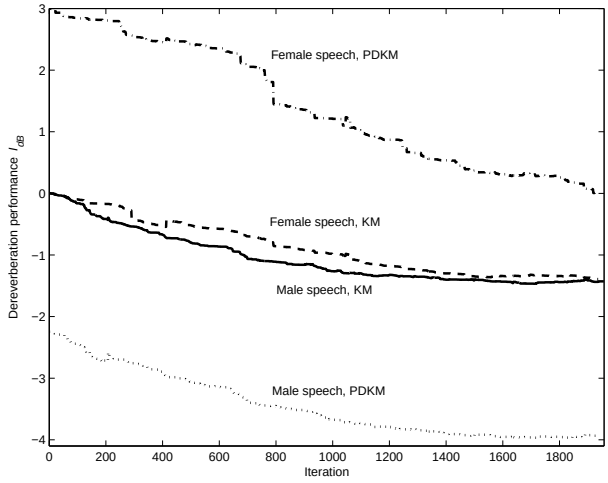

Figure 6: Dereverberation performance for male and female speech using the KM adaptive filter only and with Predictive deconvolution (PD) stage. — male speech (KM only), −− female speech (*KM* only), ··· male speech (combined *PDKM*), ·−· female speech (combined *PDKM*).

4. EFFECT OF FIRST STAGE LP PARAMETERS ON PERFORMANCE

The KM adaptive filter and combined PDKM processes were applied to female speeech using the same parameters as above. It can be seen from Figure 6 that with the use of the predictive deconvolution stage the result is actually inferior to that achieved using KM alone. It was found that the performance of the algorithm was affected by the first LP residual extraction order P_{speech_1} (labelled *short LP 1* in Figure 5) – the dereverberation performance for the predictive deconvolution stage was measured for values of $P_{speech_1} \in \{10, 20, \cdots, 100\}$ (an increased frame size of 512 samples was used). The result is plotted for male and female speech in Figure 7. From this it can be seen that the PD performance is improved with increased predictor order, significanlty so for

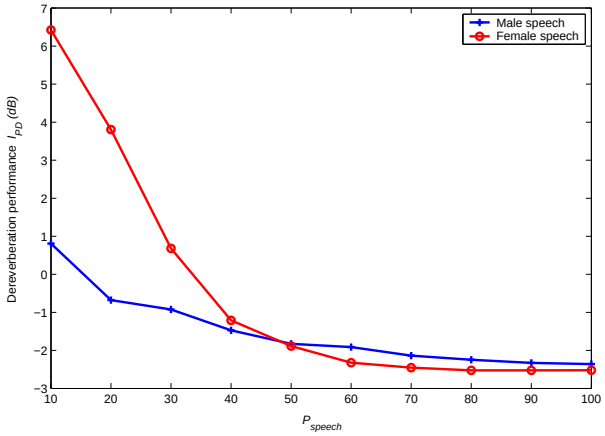

Figure 7: PD performance against pole order for male and female speech.

female speech. The optimum order was found to be approximately $P_{speech} = 80$. When the KM stage is subsequently applied as in Figure 5 (with $P_{speech_1} = 80$ and $P_{speech_2} = 20$) and significant improved is achieved. Table 1 shows the reverberant energy reduction achieved using predictive deconvolution, kurtosis maximization and the combined PD-KM approach for male and female speech with this prediction order.

	P_{speech_1}	I_{KM}	I_{PD}	I_{PDKM}
Male	20	-1.4	-2.2	-3.9
Female	20	-1.4	3.0	0.0
Male	80	-1.5	-2.3	-4.1
Female	80	-1.4	-2.5	-4.0

Table 1: Change in reverberatant energy (dB) achieved for P_{speech_1}.

A spectrogram of clean, reverberated and PDKM dereverberated male speech is shown in Figure 8. This clearly shows the improved resolution of harmonics for dereverberated speech although there is some loss in temporal definition due to the acausal nature of the KM filter response.

5. CONCLUSIONS

In this paper a blind dereverberation technique was proposed to improve upon the performance of the kurtosis maximizing subband adaptive filtering algorithm [3]. The mixed phase model of the room transfer function was exploited in which a prefiltering stage to the kurtosis maximizing filter was used to estimate the inverse of the minimum phase component. This stage is a one time FIR filter whose coefficients are found using a high order predictive deconvolution on sequences of LP speech residual. The stability behaviour of the KM algorithm was investigated and optimum values for the update step step size μ and moment smoothing constant β were determined. It was then found that, for male speech, an improved reduction in reverberant energy could be achieved using the combined technique over KM alone.

It was then shown that the performance was dependent on the prediction order P_{speech_1} of the first short LP residual extraction and that significant improvements were gained when

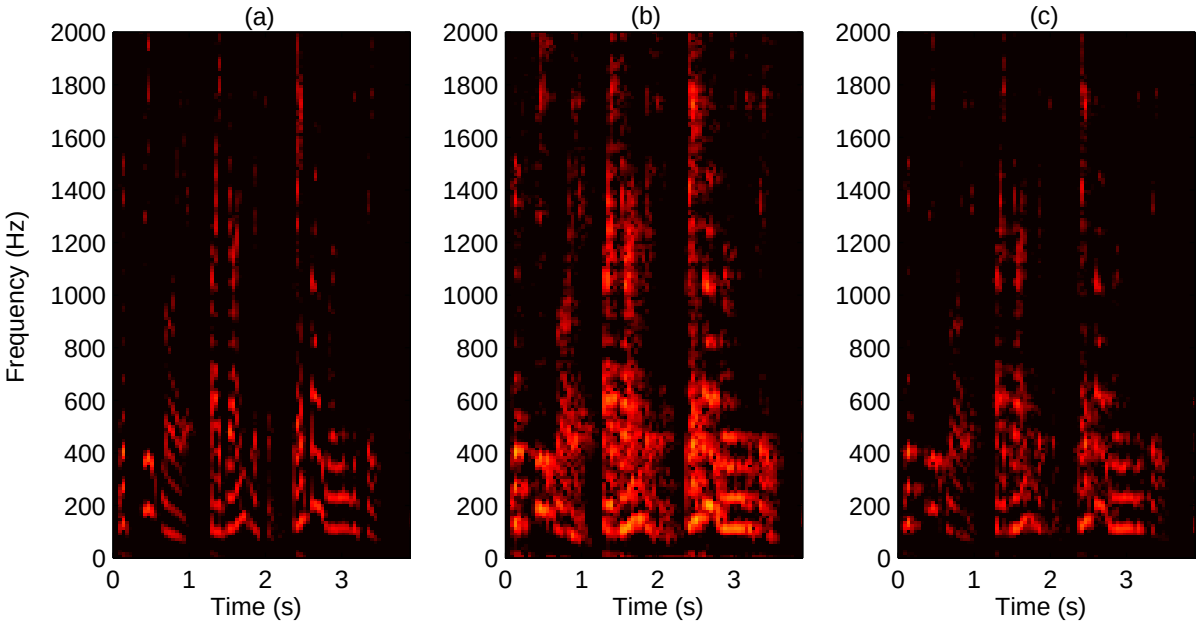

Figure 8: Spectrograms of (a) clean, (b) reverberated and (c) PDKM dereverberated male speech.

P_{speech_1} was increased, in particular for female speech. It is suggested that the stability of the KM filter is also improved by utilizing the predictive deconvolution stage.

REFERENCES

[1] Simon S. Haykin, *Adaptive filter theory*, Prentice Hall, New Jersey, 3rd edition, 1996.

[2] Lawrence R. Rabiner and Ronald W. Schafer, *Digital processing of speech signals*, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[3] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, May 2001, vol. 6, pp. 3701–3704.

[4] J. N. Mourjopoulos, *The Removal of Reverberation from Signals*, Ph.D. thesis, University of Southampton, Southampton, UK, 1984.

[5] Stephen T. Neely and Jont B. Allen, "Invertibility of a room impulse response," *Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, 1979.

[6] J. N. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, 1991.

[7] Heinrich Kuttruff, *Room Acoustics*, London: Applied Science Publishers Ltd, London, 1973.

[8] J. N. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, no. 2, pp. 217–228, 1985.

[9] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.