Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

Acoustic impulse response measurement using speech and music signals.

John Usher[1]

[1] *Barcelona Media Centre d'Innovacio — Av. Diagonal, 177, planta 9, 08018 Barcelona, Spain*

Correspondence should be addressed to John Usher (john.usher@barcelonamedia.org)

ABSTRACT

Continuous measurement of room impulse responses (RIRs) in the presence of an audience has many applications for room acoustics: in-situ loudspeaker/room equalization; teleconferencing; and for architectural acoustic diagnostics. A continuous analysis of the RIR is often preferable to a single measurement, especially with non-stationary room characteristics such as from changing atmospheric or audience conditions. This paper discusses the use of adaptive filters updated according to the NLMS algorithm for fast, continuous in-situ RIR acquisition; particularly when the input signal is music or speech. We show that the dual-channel FFT (DCFFT) method has slower convergence and is less robust to coloured signals such as music and speech. Data is presented comparing the NLMS and the DCFFT methods and we show that the adaptive filter approach provides RIRs with high accuracy and high robustness to background noise using music or speech signals.

1. INTRODUCTION

Acquisition of acoustic impulse responses in the presence of an audience has many applications for the audio engineer and acoustician [8]. In-situ IR measurement is particularly relevant for musical performances in closed spaces, where changing audience conditions can dramatically affect the reflected sound in the room and the corresponding subjective listening experience. Besides equalization of room acoustics, other applications for continuous acquisition of an RIR include teleconferencing (to recreate a near-end listening experience a far-end room); and architectural acoustic diagnostics (e.g. to measure how the real-world RIR compares with the intended RIR model in the presence of an audience).

1.1. Applications for continuous in-situ RIR acquisition

1.1.1. Room equalization

"Room equalization" is part of a family of electroacoustic processes such as "room compensation", "room correction", "reverberation reduction", "low-frequency correction", "resonance suppresion" and "modal equalization" [2, 4, 6, 18, 19]. Room EQ is typically undertaken for a number of sound quality and electroacous-

tic objectives: increasing speech ineligibility or musical fidelity; increasing the loudness of audio; and protecting the loudspeakers and electronic amplification system. The interaction of the room and loudspeaker can therefore be compensated for; namely both the low frequency response that is affected by acoustic loading [16], and reducing acoustic resonances so as to mitigate feedback. Secondly, the equalization may be motivated to affect only the response of the loudspeaker itself [24] and the room interaction effect ignored.

The present paper is concerned with applications of using a continuously updated RIR estimation for room equalization. An example for this is summarized in Fig. 1, where a loudspeaker-microphone impulse response, $\mathbf{h}$, is estimated with filter $\tilde{\mathbf{h}}$. If we wish for a maximally flat frequency response (e.g. to enhance music fidelity or speech intelligibility) then $\tilde{\mathbf{h}}$ must be inverted to generate the loudspeaker EQ filter. Before direct inversion (e.g. taking the reciprocal of the spectrum with the so-called "Kirkeby algorithm" [14] or using a time-domain approach such as the Levinson recursion algorithm), the impulse response is converted to a Minimum Phase (MP) response to avoid pre-echo time aliasing, using homomorphic decomposition [17]. The MP filter is further smoothed in both the time and complex-frequency domain to enlarge the area over which the equalization will be effective and to ensure against high-Q filters that could damage the electroacoustic system [9, 19].

1.1.2. Identification of strong early reflections

The absolute timings of early reflections in a room impulse response (RIR) can be used to solve a variety of problems relating to both subjective and physical acoustical phenomena in the corresponding environment (be it real or virtual). For instance, the delay between the direct sound arrival and the first strong reflection, i.e. the initial time-delay gap, can affect perceptual integration of the direct and reflected sound according to the precedence effect and may affect speech intelligibility or musical timbre. Physical characteristics of a space can also be determined from precise timings of reflections in an acoustic IR; a process commonly used in SONAR, seismic exploration and medical diagnosis.

The most computationally simple method to identify the Reflection Onset Timings in an RIR is from an analysis of the energy envelope. However, this first-order energy analysis can lead to false-positive reflection identification. This was shown in a recent study where a new method for identifying reflection onsets in an RIR was introduced based on a running local kurtosis analysis [22] . Such an IR analysis is shown in Fig. 2 where the effectiveness of the running-kurtosis method for identifying early reflection onsets can clearly be seen.

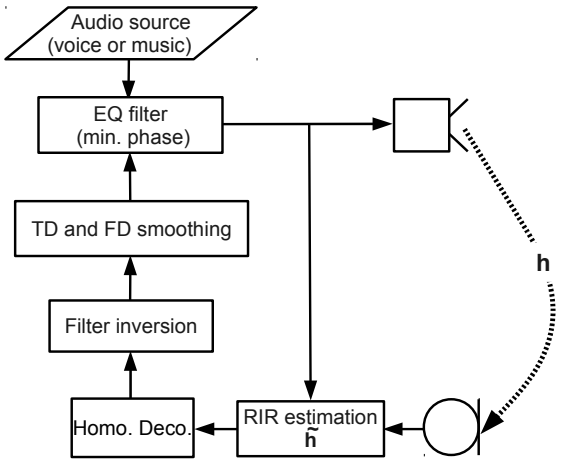


Fig. 1: Typical single-point adaptive room equalization system (filtered-x system). The loudspeaker-microphone impulse response, $\mathbf{h}$, is estimated with filter $\tilde{\mathbf{h}}$. $\tilde{\mathbf{h}}$ is inverted and homomorphic decomposition gives the minimum phase (MP) component of the adaptive filter. The MP filter is smoothed in the time and frequency domain and the result forms the EQ filter.

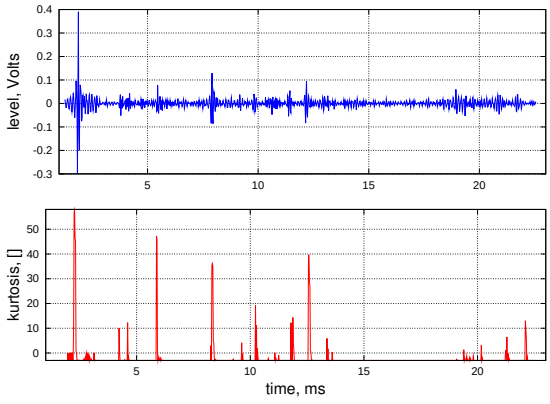


Fig. 2: RIR from a small room and the corresponding running kurtosis analysis, where the peaks coincide with the onset of early reflections.

1.2. Occupied vs non-occupied room acoustics

Berenek [1] reported that in a medium-sized concert hall, the change in the average absorption coefficient (i.e. in Sabins) with a filled vs empty audience was approximately a factor of 2 at 125, 250, and 500 Hz, and approximately a factor of 1.5 at 1 kHz, 2 kHz and 4 kHz. Accordingly, the reverberation time would be reduced by the same factor that the absorption coefficient was increased, affecting speech intelligibility and critical distance.

Hidaka et al [12] conducted an extensive investigation of 21 concert halls to investigate the change of RT in occupied vs unoccupied conditions. 6 of these halls were investigated using a conventional swept-sine technique [5] and the others were investigated using a novel technique that analysed the rate of sound decay following a musical phrase ending (a so-called "stop chord", e.g. from a church organ). However, the variation in predicted RT using the musical stop-chord method was double that using the swept sine method, and of course the RT decay curve was only valid for those frequencies where the musical excitation energy was sufficient.

Besides the increase in reflected sound energy in an occupied vs unoccupied hall, a tonal resonance is sometimes apparent in venues due to the acoustic resonance of the air space between the seat rows: the so-called seat-dip effect. The magnitude of this resonance has been reported to be up to +4 dB in the 200 Hz octave band [3], and so would likely be audible.

1.3. Dual-Channel FFT analysis to estimate an RIR

Considering that a room impulse response is defined as a filter that translates an excitation (loudspeaker) input signal into a measured (microphone) signal, the impulse response can be recovered as a complex quotient of these two spectra. However, this technique is not effective for band-limited excitation signals measured in an environment with broad-band noise, i.e. the spectral quotient becomes poorly conditioned for low signal-to-noise ratios. To overcome this, a running SNR estimate based on the cross-power spectrum or coherence function is used to weight the spectral quotient [2, 11]. A limitation of this approach is the lack of a cost function to determine optimal smoothing parameters, so if there is a sudden change in ambient noise conditions (e.g. caused by a local noise source or audience), then update of the IR may be frozen for those frequencies with low SNR, whilst the IR could meanwhile be changing (e.g. due to dynamic room boundary conditions or atmospheric changes).

2. ADAPTIVE FILTERS FOR RIR ACQUISITION

Using adaptive filters to estimate an acoustic impulse response is nothing new. Probably the most common application is echo cancellation (or echo suppression) for telecommunications. Echo cancellation is called for with a full-duplex phone-call between a far-end and near-end party where at least one party has the voice of the other radiated from a loudspeaker. This has the effect that at least one party hears their own voice as an echo when they speak. Digital acoustic echo-canceling has been implemented since the 1970's, though it was well-understood in 1966 as a patent by Kelly from Bell Labs shows [13], which was based on pioneering work by Wiener (1940's) and Widrow and Hoff (1960's). An echo-canceling system is summarized in Fig. 3: consider the far-end voice radiated with the loudspeaker, which is then detected by the near-end microphone (plus any other local sounds). An adaptive filter models the impulse response between the loudspeaker and microphone. The loudspeaker signal is then processed with this adaptive filter and subtracted from the microphone signal: removing (or reducing) the "bleed-through" of the far-end signal into the transmitted near-end signal (the "error" signal is transmitted to the far-end). The process by which the adaptive filter is adjusted to model the loudspeaker→microphone impulse response is now described.

2.1. Description of the RIR acquisition system

Fig. 3 describes the investigated RIR acquisition system. The electroacoustic impulse response, $\mathbf{h}$, between the radiated signal and detected microphone signal is approximated with the adaptive filter coefficients $\tilde{\mathbf{h}}$. The music or speech input signal is filtered with FIR filter coefficients $\tilde{\mathbf{h}}$ and subtracted from the measured microphone signal. The resulting "error signal" is used to update $\tilde{\mathbf{h}}$ so as to minimize the error signal power according to the Normalized Least Means Square (NLMS) algorithm.

The NLMS filter is well-known for its robustness to system noise and its high stability and reasonably fast convergence rates. The update of the adaptive filter is a computationally simple process, which can be accomplished in real-time on even a modest portable computing device.

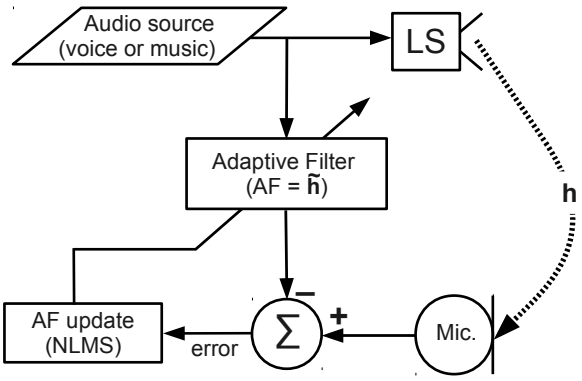


Fig. 3: Overview of the investigated RIR acquisition system. The electroacoustic impulse response, $\mathbf{h}$, between the radiated signal and detected microphone signal is approximated with the adaptive filter coefficients $\tilde{\mathbf{h}}$. The music or speech input signal is filtered with FIR filter coefficients $\tilde{\mathbf{h}}$ and subtracted from the measured microphone signal. The resulting "error signal" is used to update $\tilde{\mathbf{h}}$ so as to minimize the error signal power according to the Normalized Least Means Square (NLMS) algorithm.

A brief summary of the NLMS algorithm is now given, considering an original input signal (e.g. a music or speech signal) $x(n)$ that is filtered using the M-length adaptive filter $\tilde{\mathbf{h}}$ to give a filtered signal $y(n)$:

$$y(n) = \sum_{k=0}^{M-1} x_j(n-k)\tilde{h}_k \tag{1}$$

$$= \mathbf{x}^T(n)\tilde{\mathbf{h}}. \tag{2}$$

Where:
$\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-M+1]^T$,
$\tilde{\mathbf{h}} = [h_0, h_1, \ldots, h_{M-1}]^T$.

The microphone signal $m(n)$ is then subtracted from the filtered signal $y(n)$ to give the error signal $e(n)$:

$$e(n) = m(n) - y(n). \tag{3}$$

If the adaptive filter coefficients $\tilde{\mathbf{h}}$ match the loudspeaker-microphone electroacoustic IR exactly, then the filtered music or speech signal would exactly cancel the mic signal, giving an error signal $e(n)$ of zero energy. The error signal energy is therefore used as a *cost function* to measure the performance of the adaptive filter for modeling the electroacoustic IR, defined as the scaler J:

$$J(\tilde{\mathbf{h}}) = E\{e^2(n)\}, \tag{4}$$

where $J\{\cdot\}$ is the statistical expectation operator, i.e. the expected average of a signal. The requirement for the algorithm is to determine the operating conditions for which J attains its minimum value. This state of the adaptive filter is called the "optimal state" [10]. When a filter is in the optimal state, the rate of change in the error signal level (i.e. J) with respect to the filter coefficients $\tilde{\mathbf{h}}$ will be minimal. This rate of change (or gradient operator) is an M-length vector ∇, and applying it to the cost function J gives:

$$\nabla J(\tilde{\mathbf{h}}) = \frac{\partial J(\tilde{\mathbf{h}})}{\partial \tilde{\mathbf{h}}(n)}. \tag{5}$$

The right-hand-side of the last equations are expanded using partial derivatives in terms of the error signal $e(n)$ from Eq. 4:

$$\frac{\partial J(\tilde{\mathbf{h}})}{\partial \tilde{\mathbf{h}}(n)} = 2E\left\{ \frac{\partial e(n)}{\partial \mathbf{h}(n)} e(n) \right\} \tag{6}$$

Updating the filter vector $\tilde{\mathbf{h}}$ from time sample $(n-1)$ to time (n) is done by multiplying the negative of the

gradient operator by a constant scaler, and normalizing by the power of the input signal (i.e. so that the gain change applied to large input signals is less than for low signals) and the filter update (i.e. the steepest descent gradient algorithm) is:

$$\tilde{\mathbf{h}}(n) = \tilde{\mathbf{h}}(n-1) + \frac{\mu}{\delta + \mathbf{x}^T(n)\mathbf{x}(n)}\mathbf{x}(n)e(n) \qquad (7)$$

where the step-size is bound
$0 < \mu < 2$
and δ is a normalizing or regularization constant to ensure against computational errors when the power estimate of the input signal is too low.

Besides the increase in computational efficiency of implementing the filter-update and signal filtering in the frequency domain (requiring 5 FFTs per iteration; i.e. for every M input samples), the performance of the frequency-domain and time domain NLMS algorithm are equivalent [21]. In the present study, the overlap-save technique [20] was used with an overlap factor of four. For the filter update, the time-domain constraint (to ensure against "wrap-around" errors when M is less than the length of the actual RIR) can be affected so as to weight later coefficients less than early ones; a modification known as the "exponential step" (ES) algorithm [15].

2.2. System parameters

2.2.1. Filter length

The adaptive filter coefficients $\tilde{\mathbf{h}}(n)$ can be inspected to investigate how well they approximate the acoustic impulse response between the sound source and microphone. If the input signal to the loudspeaker is from a microphone in the same room, then feedback will occur (so-called *regeneration*). Such regeneration will be manifested as a peak in the IR at a time equal to the direct sound-path time between the loudspeaker and microphone (i.e. the time it takes sound to travel between the microphone and loudspeaker). However, if the sound-path time between the loudspeaker and microphone is greater than the length of the adaptive filter, then regeneration effects will not be modeled.

2.2.2. Step-size and regularization

A parallel multi-filter approach can be implemented whereby multiple simultaneous filters run with different step-size (mu) and regularization (δ) parameters [23]. This prevents erroneous impulse response estimates following a sudden change in the ambient noise (in echo-cancellation applications, this phenomenon is called "double-talk") or changes in the impulse response (e.g. from moving objects in an auditorium). With the multi-filter approach the "best" step-size and regularization combination can be selected, which may be such that the update is frozen between iterations (i.e. a step-size of 0 is selected).

2.3. Use of adaptive whitening filters

When "coloured" music or speech audio signals are used to adapt the filter, the Rate of Convergence (RoC) is slower and the filter response is only determined at corresponding frequencies. However, "whitening" algorithms that introduce spectral distortion to the input signal can both increase RoC and extend the spectral range over which the adaptive filter is characterized.

Two whitening approaches were considered: partial rectification and a more complex adaptive system using the (normalized) residuals from a running LPC analysis of the two input signals [7]. Simulations were conducted offline with a previously obtained small room RIR and music, speech and noise input signals.

RoC was hardly affected using the partial rectification method. However, using the LPC adaptive whitening filter we found a halving in the convergence time for speech signals in high SNR conditions. In fact, even using a pure sine-wave input the adaptive filter converged to a near-perfect reconstruction of the original RIR. Unfortunately, the limitations of the LPC whitening method were revealed as the SNR decreased, and at even modest noise levels (e.g. 30 dB SNR and lower) the obtained adaptive filter state was dramatically worse than the case when no whitening filter was used. As such, use of a whitening filter to improve RIR estimation for coloured signals was not found to be beneficial for typical SNR conditions.

3. EMPIRICAL MEASUREMENTS: USING SPEECH AND MUSIC TO MEASURE THE RIR WITH AN ADAPTIVE FILTER

3.1. Experiment configuration and method

We investigated how radiating music and speech audio signals into a room can be used to measure the IR between the loudspeaker and the microphone, using an adaptive filter updated according to the NLMS filter, as described in Fig. 3. By way of an experimental control, we first measured the impulse response between the loudspeaker and microphone using the standard exponentially swept-sine technique [5] with a 2 second swept-sine, and compared our obtained adaptive filter with this reference measurement. The comparison of the adaptive filter with the reference IR can be done in both the time domain and frequency domain, so we will show both.

Using the swept-sine method to acquire the reference IR is non-ideal because distortion artifacts introduced

by the loudspeaker are windowed-out. However, high-quality loudspeakers were used in the experiment (Genelec model 8020) and the reproduction levels were not high (approx. 80–85 dB) so distortion is expected to be minimal. This was confirmed by analysis of the pre-ringing components in the deconvolved IR where these distortion components would appear (i.e. that part of the IR before the main peak), which had very low energy.

We thought it would be elegant to use small radio "lavalier (lav) microphones" to measure an IR in a filled auditorium (e.g. they could be hidden near the maestro of an orchestra or near a front-of-house mixing desk), so high quality lav mics were used for all the data shown here (model C417L by AKG, with a flat frequency response from 50 Hz–5 kHz with a +5 dB boost at 10 kHz and -3 dB response at 20 kHz). The room used for the measurement was quite small (approx. 10 m^3), though directly coupled to a much larger volume through a corridor. The loudspeaker configuration was for a conventional 5.1 system, with the microphone at the central sweet-spot 2 m away. An analysis of the time decay of the impulse response can be seen in Fig. 10.

Speech, music and broad-band white noise signals were radiated from the loudspeaker. The speech segment was a spoken female voice, recorded anechoically ("In language, infinitely many words can be written with a small set of letters."); and the music segment was the first part of the pop song "Back to black" by Amy Winehouse. Spectra of the speech and music are shown in Fig. 11. The audio repeated for approx. 20 seconds, though the filter convergence was typically in 1–10 seconds, as is shown in the Rate of Convergence curves in Fig. 4.

Besides the three source types, a fourth condition was also investigated when live voice was spoken at the location of the (inactive) loudspeaker. For this fourth configuration, the two inputs to the adaptive filter system were:

1. The signal from a lav microphone attached to a live human speaker.
2. The signal from a distant lav microphone located 2 m away.

Such a condition could be taken as using a first near microphone of a talker at a conference and using a second distant microphone within the room: thereby estimating the impulse response between the two microphone locations.

The filter length was 2048 samples (approx. 40 ms at 48 kHz fs), which was chosen because it is primarily the early reflections which contribute to the IR spectrum, and the early reflection stage of the IR is approximately 50 ms for the room investigated (after 50 ms, the IR was stochastic noise). Each data window was overlapped by a factor of 4 for increased rate of convergence. The step size was fixed at 0.1 and the regularization coefficient was 5.

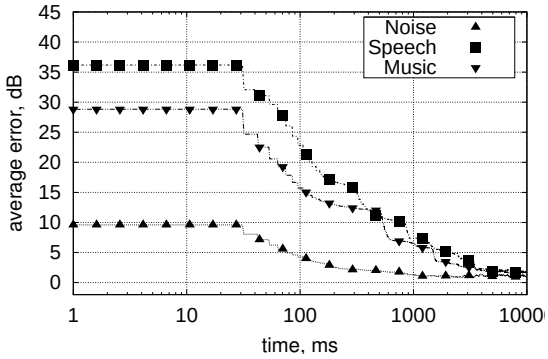


Fig. 4: Rate of Convergence for adaptive filter with noise; music and voice signals. Error is the difference between average adaptive filter level compared with the reference IR (i.e. determined using a swept sine), averaged from 80 Hz–16 kHz.

A second experiment was conducted to investigate the robustness of the adaptive filter system to different ambient signal-to-noise ratios. To affect the SNR, 6 uncorrelated channels of broad-band noise were reproduced using 6 loudspeakers spaced at 60° around the measurement location (the source signal was radiated at a location of one of the noise loudspeakers). Three SNRs were investigated: 25 dB (this was with no noise source active, and just ambient background noise in the measurement environment); 12 dB and 0 dB.

3.2. Results

Time-domain plots of the reference IR and the adaptive filter response 10 seconds from initialization are shown in Fig. 5, for the 3 different loudspeaker input signals: white noise; music and speech. A fourth condition was also investigated, when live voice was spoken at the location of the (inactive) loudspeaker and the two inputs to the adaptive filter system were:

1. The signal from a lav microphone attached to a live human speaker.
2. The signal from a distant lav microphone located 2 m away.

The corresponding frequency-domain spectra of the adaptive filters are given in Fig. 6, which were smoothed using $1/3^{rd}$ octave bands for visual clarity.

3.3. Discussion

Comparing the reference IR spectra with the adaptive filter spectra in Fig. 6, we can clearly see how closely matched the adaptive filter coefficients $\tilde{\mathbf{h}}$ are to the acoustic IR $\mathbf{h}$. The music and speech spectra do not match the reference IR at high frequencies (>10 kHz), but this is not surprising considering the low energy of the source signal here (the music and speech signal spectra are shown in Fig. 11). Likewise, at low-frequencies (<80 Hz), the speech adaptive filter spectrum does not match the reference IR, due to the band-limited content. Interestingly, when the noise signal was used, although there was near perfect alignment with the reference IR above 100 Hz (within 0.5 dB up to 22 kHz) there was noticeable low-frequency deviation. This may have been to high level low-frequency noise in the measurement (e.g. there is a distinct 50 Hz peak in the narrow-band FFT, probably from electrical noise interference with the radio microphones).

Looking at the adaptive filter generated using the live spoken voice, it is clear that both the frequency response differs significantly from the reference IR, though it follows the same general curve in terms of peaks and dips. From the time-domain adaptive filter responses the early reflections can be identified at 2.4; 4; 6.5 and 12 ms after the direct sound peak, which could be automatically detected using the running-kurtosis method [22].

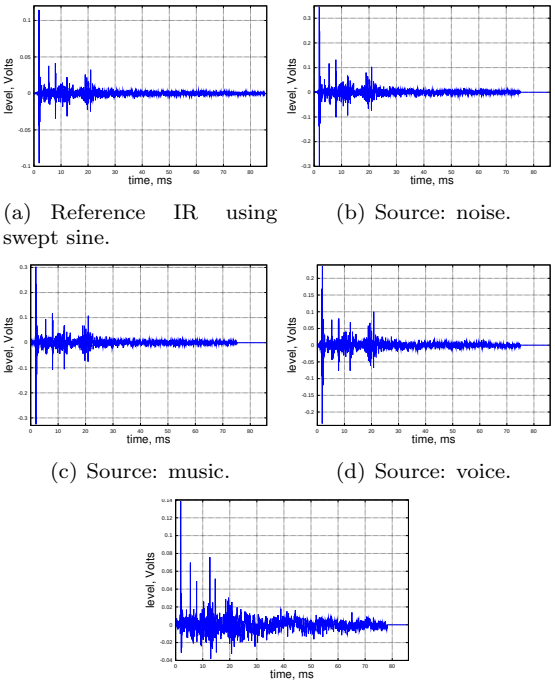
4. COMPARISON WITH DUAL-CHANNEL FFT METHOD

Following a recommendation by an anonymous AES reviewer, the NLMS adaptive filter system was compared with the professional room equalization system called "SMAART"[1] which the reviewer esteemed to be the "state of the art". According to the user manual, this system uses the dual-channel FFT analysis method.

4.1. Experiment configuration

The default configuration parameters were used with the DCFFT (i.e. SMAART) system. To investigate the optimum window length to use, we first analysed a recording of white noise emitted by a single loudspeaker and recorded by a single lav. microphone 2 m away: i.e. we analysed the same test signal as described in the previous section and used the data to the generated the IR.

[1]Manufactured by EAW. Version V.6, running on a Mac OSX 10.6.



(a) Reference IR using swept sine. (b) Source: noise.

(c) Source: music. (d) Source: voice.

(e) Source: live voice spoken at loudspeaker location.

Fig. 5: Adaptive filter responses for different source types. The reference IR was generated using a 5 sec. swept sine. The adaptive filter approximates the IR between the loudspeaker and the measurement microphone, except for (e) where the sound source was a live spoken voice located at the (inactive) loudspeaker. Filter state is after 10 seconds, with a SNR of approx. 25 dB.

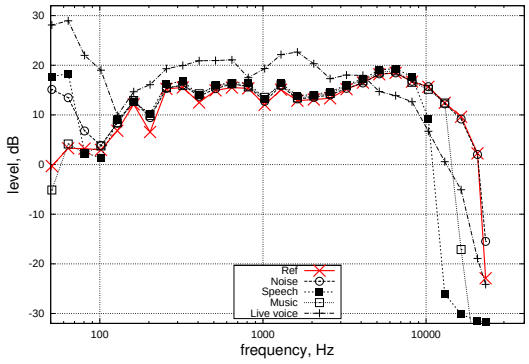
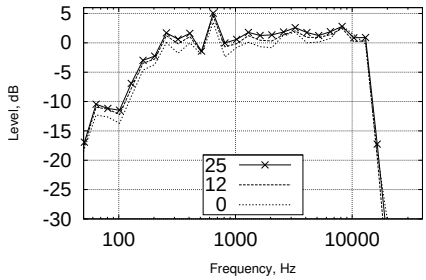
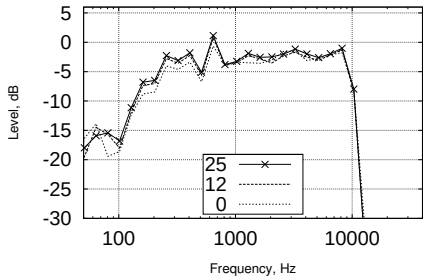


Fig. 6: $1/3^{rd}$ octave spectra of IRs from Fig. 5.



(a) Source=music. Masker=noise.



(b) Source=voice. Masker=noise.

Fig. 7: Adaptive filter $1/3^{rd}$ octave spectra for different SNRs (25, 12 and 0 dB); different sources (music or voice) radiated from 5 loudspeakers surrounding the measurement location.

The emitted test signal was fed to the reference input channel of the DCFFT system, and the recorded mic signal fed to the other "live" input signal. The playback and ADC/ DAC system were at 48 kHz, 24 bit, using professional-grade hardware. From the DCFFT system, we analyzed approximately 10 seconds of the reference and microphone signal, and adjusted the window size and number of averages. The following combinations of window sizes (in samples) and average cycles were investigated: 128k x 4 (i.e. a window size of 128000 samples and 4 averages); 256k x 2; 32k x 16; 512k x 1; 64k x 8. The window size did not significantly affect the determined IR spectra for frequencies above 100 Hz. For the following comparison of the SMAART system with the new adaptive filter system, we therefore chose a window size of 64k samples and 8 averages. To modify the SNR, white noise was added to the "live" input signal of DCFFT.[2]

In a second analysis using music and speech signals, the RIRs obtained using the DCFFT and adaptive NLMS system were compared with reference to an "ideal" RIR measured using the swept-sine technique. The measurement set-up was as follows: a single loudspeaker (Genelec model 8020) radiated 20 seconds of the test signal in a 10 m^3 room, RT60=0.2 sec.; at a presentation level of approx. 80 dB, background noise level approx. 50 dB. A microphone (Schoeps type CMC5 U with cardioid capsule) was used to record the response at a distance of 2 m. Both the loudspeaker and the microphone signals were fed to the RIR acquisition system (i.e. either the adaptive filter or SMAART system). The test signals emitted by the loudspeaker were either music or speech (see Fig. 11 for signal details). An exponentially swept sine wave (length 2 seconds) was also radiated from the loudspeaker to generate a reference IR using the swept-sine deconvolution method [5]: it is this reference IR to which the SMAART and adaptive NLMS update system were compared.

4.2. Results

Fig. 8 shows $1/3^{rd}$ octave smoothed IR spectra using SMAART for different SNRs using a white noise source.

Fig. 9 shows a comparison of RIRs generated with the adaptive NLMS update system and RIRs generated by the commercial DCFFT system using the dual-channel FFT method. RIRs were generated using empirically

[2]NB: The method of affecting SNR for the DCFFT system was different than in the previous section for the adaptive filter performance analysis, where the noise level was affected by radiated decorrelated noise in the measurement environment.

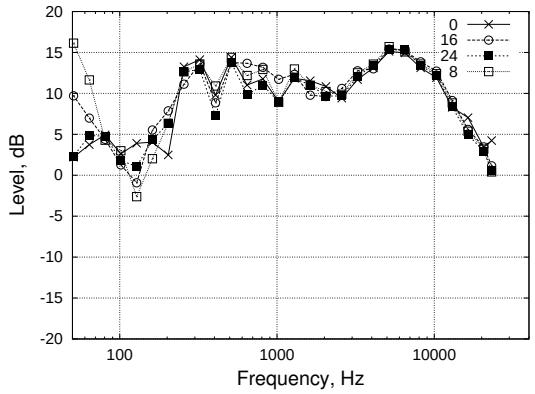


Fig. 8: 1/3rd octave smoothed IR spectra using DCFFT system for SNRs of 0, 8, 16 and 24 dB, using a white noise source. Window size = 64k samples, 8 averages.

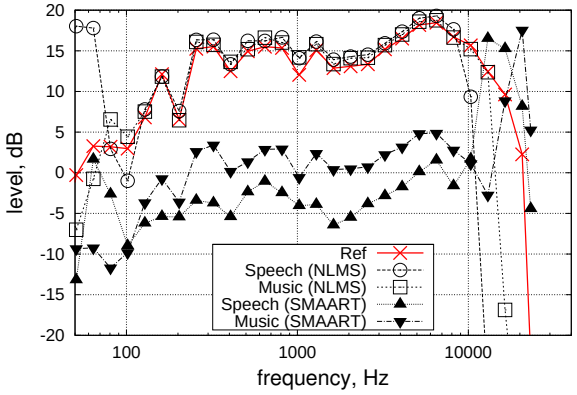


Fig. 9: Comparison of NLMS adaptive filter system with the dual-channel FFT method. Reference response made using swept-sine deconvolution. All IRs were normalized to give a variance of 0.1.

measured recordings of music and speech signals radiated from a single loudspeaker and measured with a single microphone. A reference IR is also shown generated using the exponentially swept sine wave method. All IRs were normalized to give a variance of 0.1. Spectra were averaged over 40 ms after the main peak and generated from analysis of 20 seconds of data.

4.3. Discussion

The commercial DCFFT system has comparable robustness to SNR as the adaptive NLMS system, though with the DCFFT analysis noise was used as a test-signal whereas music and speech were used for the NLMS system. For both systems, the IR spectrum differed by less than 3 dB from the highest SNR case (24 dB) to the lowest SNR (0 dB).

Looking at Fig. 9, we can see that when music and speech were used as test signals to generate the RIR the DCFFT spectra do not match the reference spectrum well. This is most pronounced in the high frequency region above 10 kHz, where the music and audio signals have low energy and the DCFFT RIRs have great high frequency noise content. This is probably as a result of the FFT inversion process where the system generated a high frequency RIR boost to account for the high frequency energy in the microphone signal. The advantage of the the cost function used by the NLMS system is demonstrated here, where we can clearly see that the estimated IR matches the reference IR within ±2 dB from 80 Hz–10 kHz.

5. CONCLUSION

Continuous acquisition of room impulse responses (RIRs) in the presence of an audience has many applications for room acoustics from loudspeaker/room equalization to forensic applications to identify the onset of early reflections. Although providing high accuracy and robustness to distortion, swept-sine and noise deconvolution are unlikely to be tolerated by an audience. Estimation of an RIR using adaptive filters, namely updated according to the NLMS algorithm, has been used for many decades in speech quality enhancement for telecommunications. The NLMS algorithm is well-suited for use with music and speech signals, due to its cost function that optimizes it's IR estimate with a minimum mean-square error criterion. A number of experiments using a single loudspeaker and microphone have shown how the adaptive system provides significant advantages over conventional FFT deconvolution methods to obtain an accurate IR estimate using speech and music signals, with fast convergence (in the order of 5–10 seconds) and high robustness to poor signal-to-noise ratios.

A further benefit of using the NLMS algorithm for RIR acquisition (though not explored in the present paper) is that the difference "error" signal can be used to capture an ambient sound signal with the injected loudspeaker signal substantially attenuated. This provides a tangible benefit for music or sports broadcasting or recording when a direct loudspeaker signal can be mixed with an audience or ambient signal: for instance, giving a much cleaner separation of the musical or speech performance and the audience applause.

6. REFERENCES

[1] L. L. Beranek. *Concert and Opera Halls: How They Sound*. Acoustical Society of America through the American Institute of Physics, Woodbury, 1996.

[2] T. Corbach, A. von dem Knesebeck, K. Dempwolf, M. Holters, P. Sorowkaand, and U. Zölzer. Automated equalization for room resonance suppression. In *Proc. of the 12th Int. Conf. on Digital Audio Effects (DAFx-09)*, 2009.

[3] W. J. Davies, T. J. Cox, and Y. W. Lam. Subjective perception of seat dip attenuation. *Acustica*, 82:787–792, 1996.

[4] S. J. Elliott and P. A. Nelson. Multiple-point equalization in a room using adaptive digital filters. *J. of the Audio Eng. Soc.*, 37:899–907, 1989.

[5] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sinetechnique. In *Proc. of the AES 108th Int. Conv.*, Paris, 2000.

[6] L. D. Fielder. Analysis of traditional and reverberation-reducing methods of room equalization. *J. of the Audio Eng. Soc.*, 51:3–26, 2003.

[7] R. Frenzel and M. E. Hennecke. Using prewhitening and stepsize control to improve the performance of the LMS algorithm for acoustic echo compensation. In *IEEE Int. Symposium on Circuits and Systems*, 1992.

[8] D. Griesinger. Beyond MLS - occupied hall measurement with FFT techniques. In *Proc. of the AES 101st Int. Conv.*, 1996.

[9] P. Hatziantoniou and J. Mourjopoulos. Errors in real-time room acoustics dereverberation. *J. of the Acous. Soc. of Am.*, 52:883–899, 2004.

[10] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, Englewood Cliffs, N. J., 4th edition, 2001.

[11] H. Herlufsen. *Dual channel FFT analysis (part I,II), Tech. Report, Brüel & Kjær Technical Review No.1, 1984, Available at http://www.bksv.com/pdf/Bv0013.pdf.*

[12] T. Hidaka, N. Nishihara, and L. L. Beranek. Relation of acoustical parameters with and without audiences in concert halls and a simple method for simulating the occupied state. *J. of the Acous. Soc. of Am.*, 109:1028–1042, 2001.

[13] J. L. Kelly. Self-adaptive echo canceller. US patent 3,500,000, 1966.

[14] O. Kirkeby, P. A. Nelson, and H. Hamada. The stereo dipole - a virtual source imaging system using two closely spaced loudspeakers. *J. of the Audio Eng. Soc.*, 46:387–395, 1998.

[15] S. Makino, Y. Kaneda, and N. Koizumi. Exponentially weighted step-size NLMS adaptive filter based on the statisticsof a room impulse response. *IEEE Trans. Acoustics, Speech and Signal Processing*, 1:101–108, 1993.

[16] A. Mäkivirta, P. Antsalo, M. Karjalainen, and V. Välimäki. Low-frequency modal equalization of loudspeaker-room responses. In *Proc. of the AES 111th Int. Conv.*, 2001.

[17] J. Mourjopoulos, P. M. Clarkson, and J. K. Hammond. A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pages 1858–1861, 1982.

[18] J. N. Mourjopoulos. Digital equalization of room acoustics. *J. of the Audio Eng. Soc.*, 42:884–900, 1994.

[19] D. Sbragion. *DRC: Digital Room Correction*. http://drc-fir.sourceforge.net/.

[20] J. Shynk. Frequency-domain and multirate adaptive filtering. *IEEE Signal Proc. Magazine*, pages 15–36, 1992.

[21] P. C. W. Sommen, P. J. VanGerwen, H. J. Kotmans, and A. J. E. M. Janssen. Convergence analysis of a frequency-domain adaptive filter with exponential power averaging and generalized window function. *IEEE Trans. on Circuits and systems*, 34(7):788–798, 1987.

[22] J. Usher. An improved method to determine the onset of reflections in an impulse response. *J. of the Acous. Soc. of Am., EL*, April, 2010.

[23] J. Usher, J. Cooperstock, and W. Woszczyk. A multi-filter approach to acoustic echo cancelation for teleconferencing. In *Proc. of the 147th Meeting of the Acoustical Society of America, New York*, 2004.

[24] R. Wilson. Equalization of loudspeaker drive units considering both on and off-axis responses. *J. of the Audio Eng. Soc.*, 39:127–139, 1991.

7. APPENDIX

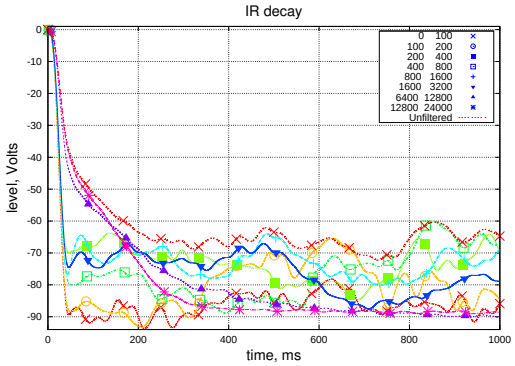


Fig. 10: IR decay envelope for the measurement room. Measured using the swept-sine technique.

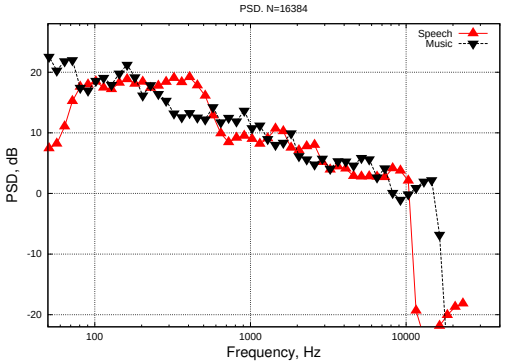


Fig. 11: Spectral profile of 15 seconds of speech and music signals used. $1/6^{th}$ octave smoothing from 340 ms Hanning windowed segments; overlapping factor=2; 15 seconds of data. The speech segment was a spoken anechoic female voice ("In language, infinitely many words can be written with a small set of letters."); and the music segment was the first part of "Back to black" by Amy Winehouse.