

Live multi-track audio recording

Joao Luiz Azevedo de Carvalho *

EE522 Project - Spring 2007 - University of Southern California

Abstract

In live multi-track audio recording, each microphone perceives sound coming from multiple sources. This degrades the overall sound quality. This work aims to resolve the “dry” sound produced from each source using a system inversion method that effectively implements room equalization and crosstalk cancelation. The proposed scheme is demonstrated in a 4-channel experiment. No audible differences were observed when comparing the original dry signals to those recovered from the live recordings. We show that the proposed method is more SNR efficient than direct equalization, and that it can be further improved with better room response measurements. Potential limitations are discussed.

1 Introduction

In music recording, each sound track is typically recorded separately, as each musician listens to a previously recorded guide track. Live recording, in which all tracks are recorded simultaneously, is commonly used for recording concerts or band practices. However, if multiple microphones at different locations are used to record different instruments and singers, each microphone may perceive sound coming from multiple sources. This degrades the overall sound quality when these tracks are combined during mixing.

This work aims to resolve the “dry” sound produced from each source in a live recording. This is achieved by using one microphone for each sound source, and the dry sounds are obtained through system inversion. This effectively implements room equalization and crosstalk cancelation.

The proposed method is demonstrated in a 4-channel experiment. The results show that the proposed method is consistently more SNR (signal-to-noise ratio) efficient than simply equalizing each channel before mixing.

*Corresponding author: jcarvalh@usc.edu

Cross-correlation evaluation reveals that the method effectively eliminates crosstalk. Simulations suggest that the method’s SNR efficiency could be considerably improved if room response measurements with higher SNR were obtained.

2 Theory

In a single-source/single-track application, the relation between the produced sound $x(t)$ and the recorded sound $y(t)$ can be modeled as a linear system $y(t) = x(t) * h(t)$, where $h(t)$ is the impulse response of the room (Figure 1). In frequency (Fourier) domain, this relation becomes a multiplication: $Y(\omega) = X(\omega) \cdot H(\omega)$. The produced “dry” sound can be obtained by measuring the room response, solving for $X(\omega)$, and inverse Fourier transforming back to time domain.

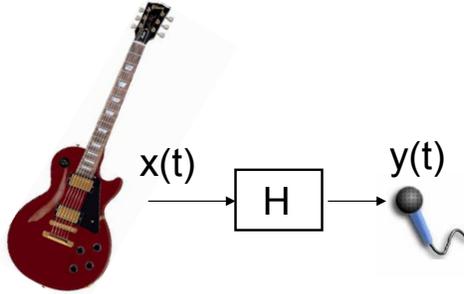


Figure 1: Linear system model of the relation between the produced sound $x(t)$ and the recorded sound $y(t)$. H is a linear system that models the room response.

In a multi-source/multi-track application where the number of microphones is equal to the number of sound sources, each microphone records the sound produced by its correspondent source (Figure 2a), but also the sound being produced by all the other sources (Figure 2b). As each source (x_i) and each microphone (j) is positioned in a different location, there is a different room response H_{ij} associated with each source-microphone pair.

The system in Figure 2b can be generalized for any number of sound sources, and can be modeled for each associated microphone as $y_j(t) = \sum_i x_i(t) * h_{ij}(t)$. In frequency domain, this system model becomes: $Y_j(\omega) = \sum_i X_i(\omega) \cdot H_{ij}(\omega)$. This can be represented in matrix form as:

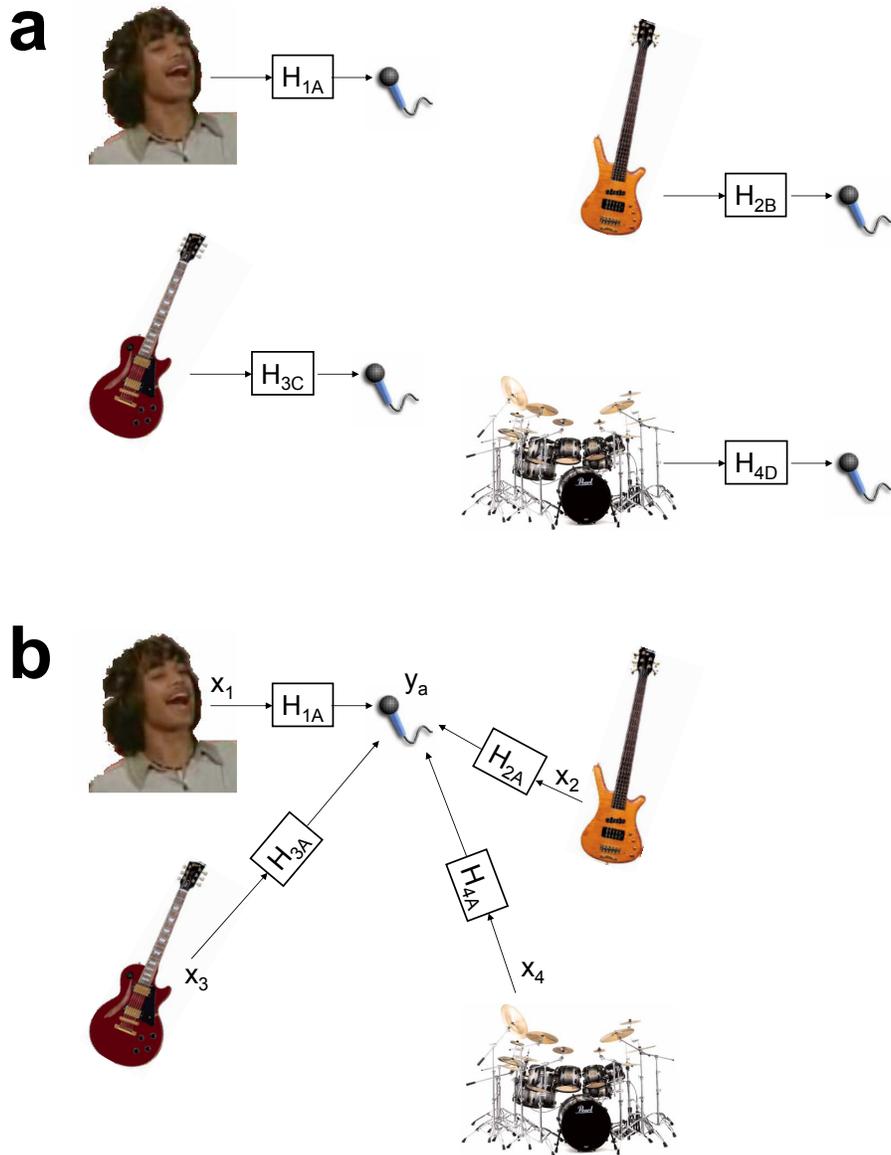


Figure 2: System model used in the proposed method. Each microphone records the sound produced by its correspondent source (a), as well as the sound being produced by all the other sources (b). A different room response H_{ij} is associated with each source-microphone pair.

$$\begin{bmatrix} Y_A \\ Y_B \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} H_{1A} & H_{2A} & \cdots & H_{nA} \\ H_{1B} & H_{2B} & \cdots & H_{nB} \\ \vdots & \vdots & \ddots & \vdots \\ H_{1N} & H_{2N} & \cdots & H_{nN} \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix},$$

or simply $Y = H \times X$.

The Y vector is formed by the Fourier transform of the set of recordings obtained from each of the microphones. The H matrix is formed from the set of frequency responses associated with each source-microphone pair. These room responses can be measured using the log-sweep technique [1], for example. The X vector contains the unknowns, as each element correspond to one of the n different sound sources in the recording.

If the number of recordings is equal to number of sound sources, this system can be solved directly by matrix inversion: $\hat{X} = H^{-1} \times Y$. The system is solved independently for each frequency component in Fourier domain. Then, each $\hat{X}_i(\omega)$ is inverse Fourier transformed back to time domain to obtain the set $\{\hat{x}_i(t)\}$, corresponding to n “dry” recordings associated with each sound source.

3 Methods

3.1 Room response measurements

We used the log-sweep technique [1] to measure the room responses $h_{ij}(t)$ associated with each source-microphone pair. For each measurement, a speaker was placed at the i -th location, and a microphone was placed at the j -th location. The following waveform was played, and simultaneously recorded:

$$a(t) = \sin \frac{2\pi f_1 T [e^{(t/T) \log(f_2/f_1)} - 1]}{\log(f_2/f_1)},$$

where the length (T) of the log-sweep signal was 3 seconds, and the frequency range covered from $f_1 = 1$ Hz to $f_2 = 22050$ Hz (Figure 3). Each measurement was repeated 10 times, and averaged, in order to achieve higher SNR.

The Fourier transform of the log-sweep signal $A(\omega)$ was obtained, as well as the Fourier transform of the averaged recordings $B_{ij}(\omega)$. The impulse responses $h_{ij}(t)$ were obtained by inverse Fourier transforming $B_{ij}(\omega)/A(\omega)$,

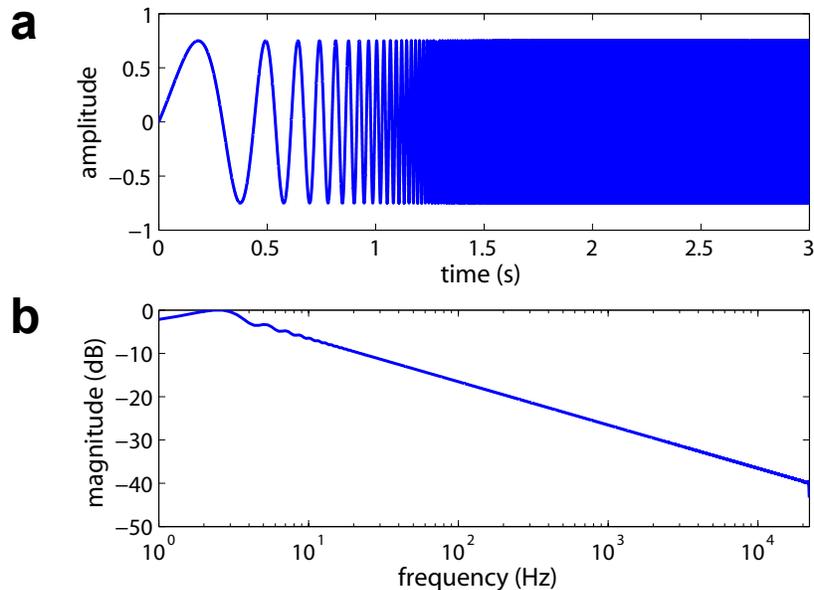


Figure 3: Log-sweep signal used to measure the room responses: (a) time domain; (b) frequency domain.

and selecting the first 372 ms. We observed that this length was enough to capture the T_{60} of the room. The frequency responses $H_{ij}(\omega)$ were obtained by Fourier transforming each impulse response.

3.2 Recordings

The proposed method was demonstrated in a 4-channel experiment. The sound recordings and room response measurements were performed as discussed next.

We used 10 second segments of four different tracks from a music CD as our set of signals $\{x_i(t)\}$. Each segment was played from a different location. A microphone was placed at each of these locations, capturing not only the direct sound from its correspondent source, but also its reflections, and the sound coming from all the other sources.

Due to hardware limitations, we could not record or produce multiple sound tracks simultaneously. Instead, we recorded each source-microphone pair separately, producing 16 recordings $y_{ij}(t)$. These recordings were synchronized, and combined to form 4 different recordings $y_j(t)$ associated with each microphone location. Due to significant background noise, each record-

ing was repeated 10 times, and averaged. The log-sweep signals, discussed in section 3.1, were recorded for each source-microphone pair immediately before each corresponding $y_{ij}(t)$ signal was obtained. The room setup is illustrated in Figure 4.

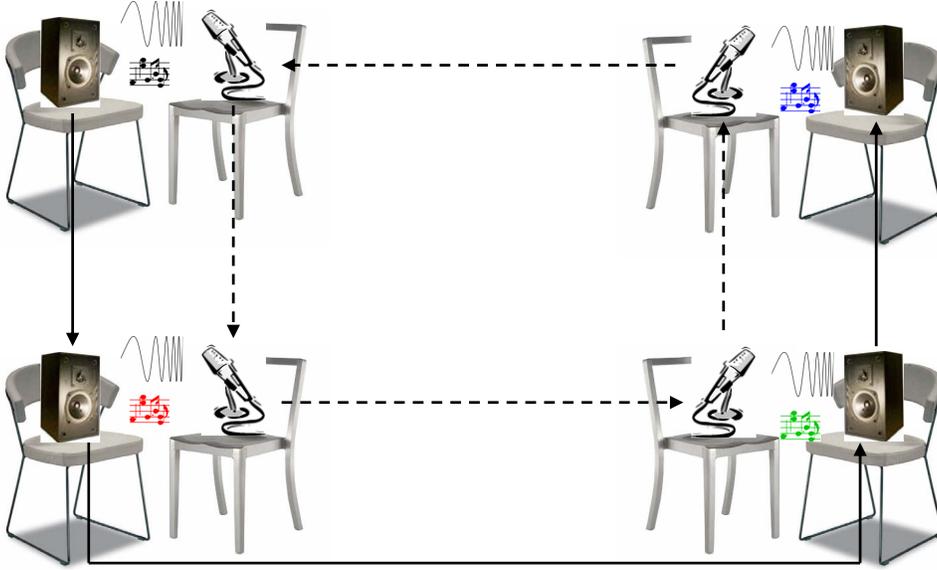


Figure 4: Room setup. The speaker was moved from locations 1 to 4, and for each location of the speaker, the microphone was also rotate from locations 1 to 4. For each speak-microphone pair of locations, a log-sweep measurement was obtained, and the sound track correspondent to the current speaker location (represented by different colors) was played and recorded.

3.3 SNR comparison

The reconstructed signals $\hat{x}_i(t)$ were evaluated in terms of SNR to the original dry signals $x_i(t)$. For comparison, we also evaluated the SNR for signals $\tilde{x}_i(t)$, obtained by simply equalizing the correspondent recorded signal $y_j(t)$ with the corresponding $h_{ij}(t)$ room response. These were obtained by inverse Fourier transforming $\tilde{X}_i(\omega) = Y_j(\omega)/H_{ij}(\omega)$.

3.4 Crosstalk cancellation evaluation

In order to evaluate the effectiveness of the proposed method in terms of crosstalk cancellation, we analyzed the cross-correlation of each of the recon-

structed signals $\hat{x}_i(t)$ to all the original dry signals $x_i(t)$. For comparison, we also calculated the cross-correlation of the dry signals to themselves, and the cross-correlation of each equalized signal $\tilde{x}_i(t)$ to the dry signals.

3.5 SNR efficiency simulation

In order to evaluate the influence of the accuracy of the room response measurements in the SNR efficiency of the proposed method, we performed the following simulation:

- The four $y_j(t)$ recordings were synthesized, rather than actually recorded, by convolving the four dry sound signals $x_i(t)$ with the corresponding room response measurements, i.e. $y_j(t) = \sum_i x_i(t) * h_{ij}(t)$.
- White noise $\eta(t)$ was added to the room responses, i.e.: $\hat{h}_{ij}(t) = h_{ij}(t) + \eta(t)$.
- Using the sets of synthesized recordings $y_j(t)$ and noisy room response measurements $\hat{h}_{ij}(t)$, we used the proposed method to obtain noisy reconstructed signals $\hat{x}_i(t)$, and used equalization to obtain equalized signals $\tilde{x}_i(t)$.
- The average SNR between $\hat{h}_{ij}(t)$ and $h_{ij}(t)$ was calculated.
- The average SNR (and standard deviation) between $\hat{x}_i(t)$ and $x_i(t)$ was calculated.
- The average SNR (and standard deviation) between $\tilde{x}_i(t)$ and $x_i(t)$ was calculated.
- The SNR of the reconstructed signals was compared to the SNR of the room response measurements.
- The amplitude of $\eta(t)$ was varied, and the experiment was repeated.

4 Results

A total of 16 log-sweep measurements $h_{ij}(t)$ were obtained. Representative room responses measured in locations near and far from the speaker are shown in Figure 5. The results indicate that components below 100 Hz and above 11 kHz are considerably attenuated. This is due to the fact that the room response measurements actually represent the system speaker-room-microphone. The speaker and/or microphone we used do not have

a flat response outside this range, and act as bandpass filters. Because of the high attenuation outside the 0.1-11 kHz frequency range (dashed lines), the filters obtained by inverting the matrix H present extremely high gains in those frequencies. Such gains would cause loss of SNR due to noise amplification. Therefore, we focused our analysis on the 0.1-11 kHz frequency range only. Components outside this range in all signals ($x_i, \hat{x}_i, \tilde{x}_i$) were nulled when evaluating the methods performance qualitatively (SNR, cross-correlation) and subjectively (listening). A wider bandwidth can be used for the reconstructed signals if equipment with more accurate frequency response is available for the impulse response measurements.

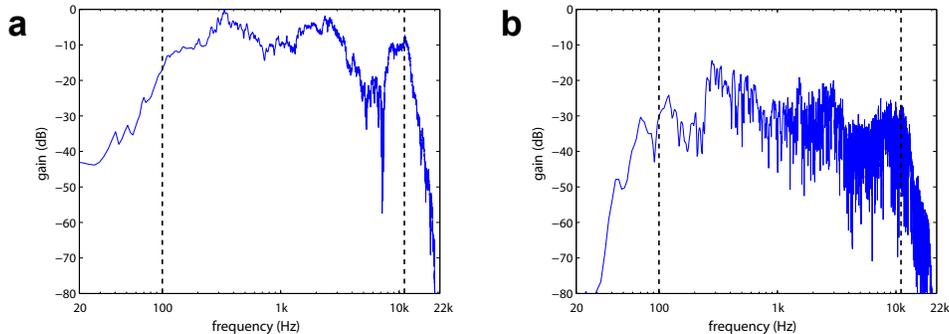


Figure 5: Representative room responses measured in locations (a) near and (b) far from the speaker. Components below 100 Hz and above 11 kHz (dashed lines) are considerably attenuated.

The SNR efficiency comparison between the proposed method and direct equalization is shown in Table 1. The results show that the proposed method is consistently more SNR efficient than equalization. In average, we observed a 5.7 dB increase in SNR using the proposed method, when compared to equalization. The SNR improvement was considerably audible. Background noise was clearly heard in the equalized signals, but was not audible in the signals recovered using the proposed method.

The results of the crosstalk cancellation evaluation are shown in Figure 6. These results show that the cross-correlation between different signals increase in the equalized results (b) when compared to the reference cross-correlations (a). This is due to the fact that each microphone captures not only the direct sound from its respective speaker (and its reflections), but also sound from all the other source (speakers in different locations). Equalization compensates the room response, effectively eliminating reflections of the correspondent sound. However, it does not address crosstalk, and back-

Table 1: SNR efficiency comparison between the proposed method and direct equalization (dB).

| | Track 1 | Track 2 | Track 3 | Track 4 | Average |
|---------------------|---------|---------|---------|---------|---------|
| Equalization | 9.4 | 6.4 | -0.4 | 8.5 | 6.0 |
| Proposed | 12.0 | 11.2 | 11.6 | 12.0 | 11.7 |
| Improvement | 2.6 | 4.8 | 12.0 | 3.5 | 5.7 |

ground noise from all the other sources is heard in the equalized signals. This background noise is considerably audible in the equalized recording. If the error waveform $x_i(t) - \tilde{x}_i(t)$ is played, it becomes clear that this background noise is composed by the sum of all other sound tracks. The results in (c) show that the proposed method effectively reduces crosstalk, and the original cross-correlations are recovered. No significant background noise was heard in the reconstructed signals. By playing the error waveform $x_i(t) - \hat{x}_i(t)$, we notice that the error consists of a considerably attenuated version of the correspondent sound track. Other sound tracks are not audible in the error waveform, indicating that crosstalk was effectively reduced.

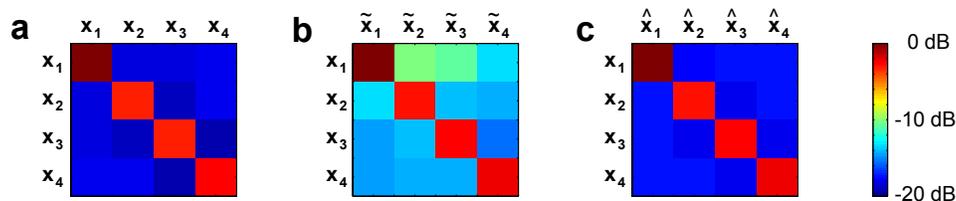


Figure 6: Crosstalk cancellation evaluation. The maximum cross-correlation value for each pair of signals is shown. (a) dry signals (reference); (b) equalization; (c) proposed method.

Figure 7 shows the results of the SNR efficiency simulation. The simulation results show that the SNR efficiency of the proposed method can be highly improved if more accurate room response measurements can be obtained. The maximum SNR for direct equalization was limited to approximately 10 dB, as crosstalk noise is not reduced using this approach. Improving the SNR of the room response measurements above 10 dB did not improve the equalization performance in this 4-channel simulation.

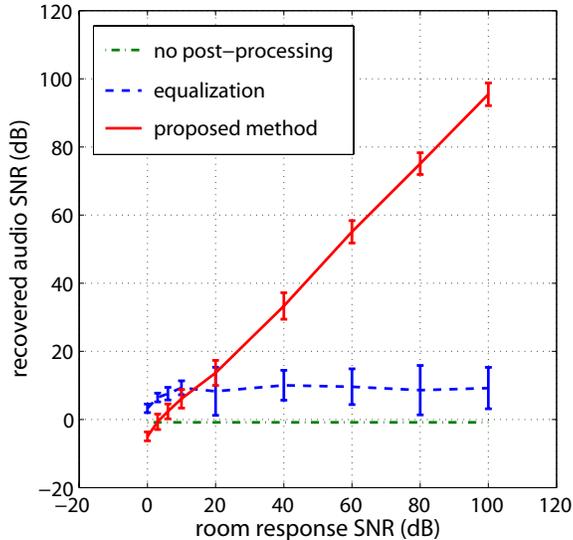


Figure 7: SNR efficiency as a function of room response measurement accuracy in a 4-channel live recording. These simulation results show that the SNR efficiency of the proposed method is linearly improved as more accurate room response measurements are used. With equalization, the maximum SNR is limited by crosstalk noise.

5 Discussion

Equalization compensates the effects of the room, and successfully resolves the dry sound in a single-track recording. However, in a multi-track recording it is not efficient. Equalization does not deal with crosstalk from the other sound sources, which appears as background noise, and degrades the SNR. The proposed method for multi-track audio recording not only achieves equalization in all tracks, but also improves the SNR by effectively reducing crosstalk. High SNR can be achieved if the room responses can be accurately measured.

One limitation of the proposed method is the requirement of multiple measurements of the room response. This might be laborious and time-consuming. Ideally, these measurements would have to be obtained with the band already inside the recording studio, and the technician performing these measurements should not be inside the room during these acquisitions. In a concert recording, the measurements should ideally be obtained with the audience already present and silent, which in practice can not be achieved.

Another limitation is the need of one microphone for each sound source. Some instruments are typically not recorded using microphones, being connected directly to the mixing/recording equipment. Also, sound sources that typically would not be recorded (e.g., monitor speakers) may need to be taken into account.

Furthermore, additional speakers might be needed for playing the log-sweep signals, because some of the sound sources may not be produced by speakers (e.g., singers, audience, acoustical instruments). Also, musicians may prefer the frequency response of their amplifiers not to be equalized. In these cases, speakers with a flat response would have to be used to play the log-sweep signals. These additional speakers would have to be placed very close to their correspondent sound sources.

The proposed method is computationally intense and should be used as a post processing stage. The content may be segmented in short blocks (e.g., 10 seconds) to reduce the computational load associated with the Fourier transforms. In this case, some overlap between these blocks could be used to avoid edge artifacts. The computational complexity associated with the matrix inversions increases quadratically with the number of audio tracks.

6 Conclusions

We addressed the issue of multi-track live recording. The proposed method resolves the dry sound produced from each source by using one microphone for each sound source, and inverting a linear system that models the recording environment. The matrix describing this system is obtained through multiple log-sweep measurements.

The proposed method was demonstrated in a 4-channel experiment. No audible differences were observed when comparing original and recorded signals. Quantitative results showed that this scheme is more SNR efficient than direct equalization. A cross-correlation analysis showed that the method effectively eliminates crosstalk. Simulation results showed that the SNR efficiency can be significantly improved if room responses can be more accurately measured. Potential limitations were discussed.

References

- [1] Farina A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Proc 110th Conv Audio Eng Soc Paris, France, 2000.