Open access • Proceedings Article • DOI:10.1109/ICASSP.2010.5495326

Latent-variable decomposition based dereverberation of monaural and multi-channel signals — Source link

Rita Singh, Bhiksha Raj, Paris Smaragdis

Institutions: Carnegie Mellon University, Disney Research, Adobe Systems

Published on: 14 Mar 2010 - International Conference on Acoustics, Speech, and Signal Processing

Topics: Spectrogram, Audio signal, Audio signal processing and Reverberation

LATENT-VARIABLE DECOMPOSITION BASED DEREVERBERATION OF MONAURAL AND MULTI-CHANNEL SIGNALS

Rita Singh

Carnegie Mellon University
Pittsburgh, PA, USA
rsingh@cs.cmu.edu

Bhiksha Raj

Disney Research
Pittsburgh, PA, USA
bhiksha@disneyresearch.com

Paris Smaragdis

Adobe Inc.
Newton, MA, USA
paris@media.mit.edu

ABSTRACT

We present an algorithm to dereverberate single- and multi-channel audio recordings. The proposed algorithm models the magnitude spectrograms of clean audio signals as histograms drawn from a multinomial process. Spectrograms of reverberated signals are obtained as histograms of draws from the PDF of the sum of two random variables, one representing the spectrogram of clean speech and the second the frequency decomposition of the room response. The spectrogram of the clean signal is computed as a maximum-likelihood estimate from the spectrogram of reverberant speech using an EM algorithm. Experimental evaluations show that the proposed algorithm is able to greatly reduce the reverberation effects in even highly reverberant signals captured in auditoria and other open spaces.

Index Terms— acoustic signal analysis

1. INTRODUCTION

Reverberation affects the quality of audio signals in most recording environments. Delayed and filtered copies of a signal from reflections off walls and other objects, interfere with the direct signal from the audio source to the listener (which might be a recording device), distorting it. While small amounts of reverberation are often tolerable and even appreciated by human listeners, longer reverberations (obtained from reflections that persist over extended periods of time) can greatly reduce the perceptual quality of the signal. In auditoria or other big recording spaces, it can often render the signal unintelligible.

A variety of techniques have been proposed in the literature to dereverberate signals. Most of them take advantage of the fact that reverberation is primarily the effect of a linear filter – the room response of the recording space – on the signal. This has led to the proposal of several *homomorphic* techniques for dereverberation (*e.g.* [1]), which take advantage of the fact that linear filters factor out as additive terms in signal cepstra. The problem with these approaches has been that the typical analysis window used to analyze the signals is usually less than 100ms long. Signal characteristics, particularly for speech, change greatly over longer time periods and the use of longer analysis windows is inappropriate. On the other hand for most typical recording environments, the reverberation time, which is characterized by their T_{60}, the time taken for the reverberations of an impulse to be attenuated by 60 decibels, usually exceeds the length of this analysis window, and can sometimes extend into several seconds. Homomorphic methods cannot account for the reverberation effects that exceed the length of the analysis window.

Variants of the above approach compute inverse filters to cancel the effect of reverberation. These methods often make various assumptions about the audio signal, such as harmonicity, independence between samples etc. to arrive at the inverse filter (*e.g.* [2], [3]). Frequently, these assumptions (e.g. harmonicity) are specific to a particular type of signal, e.g. speech. Still other methods employ other models such as codebooks and switching linear dynamic systems [4] to represent the signal for dereverberation. Once again, all of these methods suffer from one or more of the following problems: analysis windows that are shorter than the reverberation, assumptions about the underlying signal, or reliance on detailed models of the clean audio that are frequently not available. In [5], a non-negative matrix factorization based method is presented that only makes assumptions about the sparsity of the distribution of energy in spectral bands, and is similar in concept to the approach presented in this paper.

In this paper we present a new approach to dereverberation of speech signals. The technique employs the latent variable model presented in [6] to represent the process that generates the spectrogram of any sound. Reverberation is approximated as a non-negative filtering of individual spectral bands of the clean speech signal, and that process too is modeled by a latent-variable generative model. Dereverberation is achieved by estimating the parameters of this model, while imposing a minimum-entropy constraint on the process that generates the clean spectrogram of the clean speech. Experimental evaluations on real recordings in highly-reverberant and noisy environments shows that the process is able to greatly reduce the reverberation in these signals and increase their intelligibility, albeit with some minor artifacts.

One common approach to minimizing the effects of reverberation on audio signals is to capture them simultaneously with multiple microphones. A variety of array-processing techniques may then be applied to minimize the effect of the reverberation (*e.g.* [7]). However, these methods commonly require careful placement of microphones (such that the room response observed by the microphones is not significantly different) and localization of the sound source, a notoriously difficult problem in reverberant conditions. The approach we present in this paper extends easily to the dereverberation of multi-channel audio, which can deal with multi-channel signals from arbitrarily placed microphones with very different room responses without requiring localization.

The rest of the paper is arranged as follows: in Section 2 we describe the basic signal model we employ to characterize the reverberant signal. In Section 3. we outline the proposed algorithm for monaural recordings. In Section 4. we describe the multi-channel extension of the algorithm. In Section 5 we describe our experiments and finally in Section 6 we present our conclusions.

2. MODELLING REVERBERATION

Let $x[l]$ be a clean audio signal produced by some source. The sound is produced in a reverberant room. Let $h[l]$ be the room impulse response (RIR) from the sound source to the microphone recording the sound. $h[l]$ represents the reverberation in the room. The reverberant signal $y[l]$ that is actually recorded is given by

$$y[l] = x[l] \otimes h[l] = \sum_{p=0}^{L} h[p]x[l-p] \qquad (1)$$

where $\otimes$ represents the convolution operation and L is the length of the RIR and relates to the T_{60} of the room. The source signal $x[l]$ can be expressed by its Gabor representation

$$x[l] = \sum_{m} \sum_{k=0}^{N-1} X(m,k) w_s(l-mB) W_N^{k(l-mB)} \qquad (2)$$

where $W_N^k = exp(-j2\pi k/N)$, $X(m,k)$ represents the k^{th} frequency component of the m^{th} spectral vector in the *short-time Fourier transform* (STFT) of $x[l]$. N represents the length of the analysis window used to derive the STFT, and B is the number of samples by which adjacent analysis frames shift. $X(m,k)$ is given by

$$X(m,k) = \sum_{l} x[l] w_a(l-mB) W_N^{-k(l-mB)} \qquad (3)$$

$w_a[l]$ and $w_s[l]$ are the biorthogonal analysis and synthesis windows for the STFT respectively.

The STFT of the *reverberated* signal, $y[l]$ can be approximated by the convolution

$$Y(n,k) \approx \sum_{l=0}^{L_H} X(n-l,k)H(l,k) = X(n,k) \otimes_n H(n,k) \quad (4)$$

where $L_H = \lfloor (L+N-1)/B \rfloor$ and $H(l,k) = W_N^{k(N-1)} \sum_{n=0}^{2N-2} h[mB+n-N+1]w_h[N-n-1]W_N^{-kn}$ is the STFT of the RIR, $h[l]$. $w_h[l]$ is the convolution of $w_a[l]$ and $w_s[l]$. $\otimes_n$ represents a convolution operation along n.

3. STATISTICAL MODEL FOR REVERBERANT SPECTROGRAMS

We employ a model proposed in [6] to represent the spectrograms. According to this model, the magnitude spectrogram of any sound is actually a histogram of draws from a bivariate distribution over time and frequency indices. Although this is an artificial construct and does not represent any physical generating process, it has been demonstrated to be highly effective for various problems such as signal separation, component discovery and learning of overcomplete codebooks of bases [8]. The magnitude spectrogram $|S(n,k)|$ of the clean audio signal is assumed to be a histogram drawn from a bivariate distribution $P_S(n,k)$ over the discrete random variables n and k. Similarly $|H(n,k)|$ is assumed to be a histogram drawn from the distribution $P_H(n|k)$. $|Y(n,k)|$ is assumed to have been generated by a process that first draws a the tuple (n_1,k) from $P_S(n,k)$, then draws n_2 from $P_H(n|k)$, and finally produces (n,k) where $n = n_1 + n_2$. The generating process is illustrated in Figure 1. Using this model, we can now write

$$|Y(n,k)| \;=\; CP_S(n,k) \otimes_n P_H(n|k) + R(n,k) \qquad (5)$$

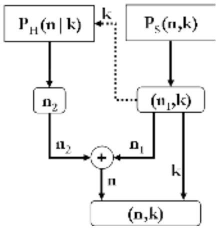

Fig. 1. Generative model for the histogram representing the magnitude spectrogram of reverberant speech.

C is a scaling constant and $R(n,k)$ captures the natural variations from the mean that occur in any drawing process. $\otimes_n$ represents a convolution operation along n. In practice, we can relate the terms in the above equation to those in Equation 4 as $P_S(n,k) \propto |S(n,k)|$ and $P_H(n|k) \propto |H(n,k)|$. $R(n,k)$ represents the correction term to be factored in to account for the fact that the magnitude of the sum of complex numbers is not equal to the sum of their magnitudes. Alternately, we can write

$$|Y(n,k)| \sim P_S(n,k) \otimes_n P_H(n|k) \qquad (6)$$

4. SINGLE CHANNEL ALGORITHM FOR DEREVERBERATION

Based on Equation 6 we cast the problem of dereverberating the audio signal as follows: given only the STFT $Y(n,k)$ of the reverberant speech we must estimate $P_S(n,k)$ and $P_H(n|k)$. The magnitude of the STFT of the underlying clean speech is given by $|S(n,k)| = CP_S(n,k)$. The scaling factor C ensures that the sum of spectral magnitudes in the reverberated signal is the same as that in the dereverberated one. The clean audio signal is obtained from $|S(n,k)|$ by "stealing" the phase of $Y(n,k)$ to obtain a complex STFT, which may be inverted to obtain a time-domain signal.

It is clear from Equation 6 that the problem is under specified. To compensate for this, we must provide some kind of *a priori* model for $P_S(n,k)$. Our prior is based on the observation that the magnitude spectra of most sounds is very sparse – at any time there are only a few frequency components with high energy. Alternately viewed, the entropy of $P_S(n,k)$ is low. Hence, we specify that the *a priori* probability distribution of $P_S(n,k)$ is given by $P(P_S) \propto exp(-\alpha H(P_S))$, where $H(P_S)$ is the entropy of $P_S(n,k)$, and α is a weighting term. This prior imputes higher *a priori* probability to distributions ($P_S(n,k)$) that have lower entropy.

We can now derive the update rules for the estimation of $P_S(n,k)$ and $P_H(n|k)$ using the Expectation Maximization (EM) algorithm. The update rule for $P_H(n|k)$ is given by

$$P(m|n,k) = \frac{P_S(m,k)P_H(n-m|k)}{\sum_{m'} P_S(m',k)P_H(n-m'|k)} \qquad (7)$$

$$P_H(n|k) = C_1 \sum_{m} |Y(n+m,k)|P(n|m,k) \qquad (8)$$

C_1 is a normalizing constant that ensures that $P_H(n|k)$ sums to 1.0.

To obtain the update for $P_S(n,k)$ we have

$$q(n,k) = \sum_m |Y(m,k)|P(n|m,k) \tag{9}$$

$$\frac{q(n,k)}{P_S(n,k)} + \alpha + \alpha \log P_S(n,k) + \rho = 0 \tag{10}$$

$$P_S(n,k) = \frac{-q(n,k)/\alpha}{\mathcal{W}(-q(n,k)e^{1+\rho/\alpha}/\alpha)} \tag{11}$$

ρ is a lagrange multiplier. $\mathcal{W}(\theta)$ is Lambert's W function. $P_S(n,k)$ is obtained through fixed point iterations of Equations 10 and 11. Typically 2-3 iterations are sufficient.

To initialize the algorithm we initially set $P_S(n,k) \propto |Y(n,k)|$ and $P_H(n|k) = 1/n, 0 \le n < N$, where N is the assumed length of the STFT of the RIR. Typically we set N to be equal to half of the T_{60}, which in turn can be estimated using algorithms such as [9].

5. MULTICHANNEL EXTENSION

Multichannel recordings have multiple recordings of the form

$$y_j[l] = x[l] \otimes h[l] = \sum_{p=0}^{L} h_j[p]x[l-p] \tag{12}$$

where $y_j[l]$ is the signal captured by the j^{th} microphone and $h_j[l]$ is the room response observed by the j^{th} microphone. The STFT of $y_j[l]$ is given by

$$Y_j(n,k) \approx X(n,k) \otimes_n H_j(n,k) \tag{13}$$

where $Y_j(n,k)$ is the STFT of $y_j[l]$ and $H_j(k,n)$ is the STFT of $h_j[l]$. Note that Equations 12 and 13 assume that the RIRs observed in the different channels are entirely different; only the underlying (and unobserved) clean audio is identical for all channels.

The corresponding statistical model is

$$|Y_j(n,k)| \sim P_S(n,k) \otimes_n P_H^j(n|k) \tag{14}$$

$P_H^j(n|k)$ is the bivariate multinomial which models $|H_j(n,k)|$. Equation 14 states that $|Y_j(n,k)|$ is the histogram of observations obtained by drawing (n_1,k) from $P_S(n,k)$, n_2 from $P_H^j(n|k)$ and forming the final observation as $(n,k) = (n_1+n_2,k)$.

Given $Y_j(n,k)$ for all channels, we must now estimate $P_S(n,k)$. The update rules for this estimation can be derived as before using EM, and are given by

$$P_j(m|n,k) = \frac{P_S(m,k)P_H^j(n-m|k)}{\sum_{m'} P_S(m',k)P_H^j(n-m'|k)} \tag{15}$$

$$P_H^j(n|k) = C_1 \sum_m |Y_j(n+m,k)|P_j(n|m,k) \tag{16}$$

C_1 is a normalizing constant as before. The updates for $P_S(n,k)$ now become

$$q(n,k) = \sum_j \sum_m |Y_j(m,k)|P_j(n|m,k) \tag{17}$$

$$\frac{q(n,k)}{P_S(n,k)} + \alpha + \alpha \log P_S(n,k) + \rho = 0 \tag{18}$$

$$P_S(n,k) = \frac{-q(n,k)/\alpha}{\mathcal{W}(-q(n,k)e^{1+\rho/\alpha}/\alpha)} \tag{19}$$

Note that this is identical to the update rules for monaural signals, with the difference that q is now obtained by averaging over all channels. Note that the fact that the same $P_S(n,k)$ is assumed for all channels reduces the degree of underspecification of Equation 14; nevertheless the overall model remains underspecified and the entropic prior must still be employed. As before, the dereverberated spectrogram is obtained as $S(n,k) = CP_S(n,k)$. C is now set to equalize the total of all magnitude spectral values of $S(n,k)$ and one of the channels: $Y_j(n,k)$.

6. RESULTS

A number of different experiments were run to evaluate the proposed algorithm.

Monaural Dereverberation: In the first experiment clean speech signals were reverberated with an artificially generated room response with times (T_{60}) between 0 and 2 seconds. The room response was obtained with the image method for a room of dimensions 3m x 4m x 5m. The synthesized signals were all monaural.

The signals were then dereverberated using the proposed method. The signals were analyzed with an STFT that employed windows that were 64ms wide. Adjacent windows overlapped by 48ms. In all cases, the reverberation time was assumed to be known (as mentioned earlier, this is not a bad assumption; reverb times can be estimated using techniques such as those in [9]). The width of $H(k,m)$, the time-spectral representation of the room impulse response, was assumed to be half the known RIR length. In order to ensure that all frequencies contribute equally to the overall estimation, the STFT of the signals were first "balanced" by normalizing every time-frequency element as follows: $|\hat{Y}(n,k))| = |Y(n,k)|/\frac{1}{N}\sum_m |Y(m,k)|$.

Table 1 shows the estimated SNR improvements obtained from dereverberation. The SNR was computed by comparing the dereverberated signal to the original clean signal, and characterizing all differences as noise. In order to compute the SNR the two signals first had to be aligned, since the room reverberation and subsequent deconvolution introduces a shift. The reconstructed signal also had to be scaled to have the same RMS value as the original clean signal.

T_{60}	0	0.2	0.5	1.0	2.0
SNR(dB)	-2.4	3.2	3.3	2.8	1.6

Table 1. SNR improvements as a function of reverb time.

We note that the improvement in SNR is superior to those obtained with technique previously reported by Kameoka *et. al.* in [5]. The improvement in SNR is observed to reduce with increasing reverb time. SNR measurements, however, are highly suspect here. The RIR introduces an attenuation that must be normalized out by scaling the signal after dereverberation. Different scaling factors can result in different SNR estimates. The perceived improvement in signal quality was typically much greater than that indicated by the SNRs in Table 1. At reverb times over 0.5 second, it was found advantageous to dereverberate the data repeatedly assuming an RIR time of 0.5 seconds each time until the desired RIR was obtained, instead of only once with the true RIR. For instance, at a T_{60} time of 2.0 seconds, repeated application of the algorithm with a dereverb time of 0.5 seconds resulted in an additional improvement of over 2.0 dB.

More realistic results are obtained by evaluating the data on *real* reverberant recordings. In a second experiment, we dereverberated two real recordings of highly reverberant speech. The first was a recording of an arabic preacher delivering his discourse in an open space. The recording was captured by a microphone mounted at a distance of several meters. Figure 2 show the spectrogram of the signal before and after dereverberation. For this experiment, since the reverb time was not known, a reverberation time of 6 seconds was assumed. The second was a recording of a famous Indian play, "Adrak ke Panje", which holds the Guinness record for the longest running play (having run significantly longer than "The Mousetrap"). The only available recordings for this play, however, were captured in a highly reverberant auditorium. Figure 3 show the spectrogram of a sample of this recording, and the dereverberated signal obtained. Once again, a reverb time of 6 seconds was employed. In both examples we observe that the dereverberation algorithm significantly reduces the smearing of the spectrogram that is caused by reverberation. Perceptually, we also observe musical noise, introduced by the enforcement of sparsity which floors some time-frequency components. Samples of the dereverberated audio may be heard at http://www.cs.cmu.edu/b̃hiksha/audio/dereverb

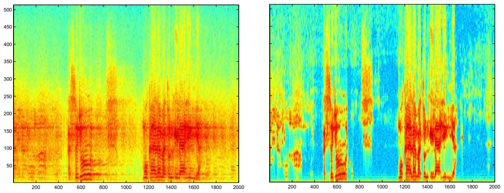

Fig. 2. Left: Spectrogram of two seconds of highly reverberated speech from our arabic example. Right: Dereverberated version of the same sample.

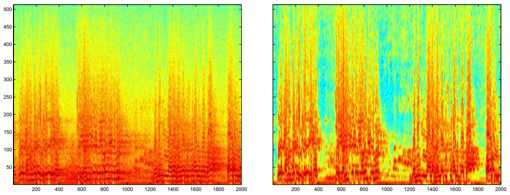

Fig. 3. Left: Spectrogram of reverberant auditorium sample. Right: Dereverberated version of the same signal.

Multi-channel audio: The multichannel version of the algorithm was evaluated using synthetic multi-channel data. 11-channel recordings were obtained by simulating a linear microphone array with 5cm spacing, in a 12m x 5m x 4m room. The array is centered 2m from the far wall of the room, and the speaker is 8m from the array and slightly off center (2.6m and 2.4m from the two walls, respectively). The image method was used to generate the room response.

A T_{60} time of 2.0 seconds was assumed for the room. Table 2 shows the SNR improvement obtained by dereverberation. In our experiments below we chose one of the microphones in the center as the primary microphone (for the 1-channel case), and expanded symmetrically outward to increase the size of the array.

N. channels	1	2	4	8
SNR(dB)	1.6	1.8	2.4	2.2

Table 2. SNR improvements as a function of the number of microphones.

We note that the improvement in SNR from the dereverberation increases as the number of channels increases. In particular, the SNR improvement with 8 microphones was comparable to that obtained with a delay and sum beamformer on the same data. However, unlike delay and sum, our algorithm does not require the microphones to be arranged as a calibrated array, and the room responses observed by the various microphones can be considerably different.

As stated earlier, the SNR measurements in the above tables are highly imperfect. The true quality of the results produced by the algorithm is best judged from the audio samples at http://www.cs.cmu.edu/b̃hiksha/audio/dereverb

7. DISCUSSION

The dereverberation algorithm is observed to very effective at eliminating the smearing that occurs in the spectrogram of a signal as a result of reverberation. Perceptually, too, the dereverberated signals sound significantly more crisp than the reverberant signals. Informal tests show that they are in fact also more intelligible, particularly for the highly reverberated audio such as that in our arabic and auditorium samples. The quality of the signal still leaves something to be desired though – the dereverberated signals have significant musical noise. We are currently developing algorithms that compose Wiener filters from smoothed versions of the dereverberated spectrograms to eliminate this problem. Also, the imposition of sparsity on the spectrograms of the dereverberated signal make them inappropriate for speech recognition. We believe that the Wiener filter framework will result in more natural spectrograms that can result in significant improvements in recognition accuracy for reverberated speech. We are also working towards developing alternatives to the entropic prior to obtain more natural sounding speech.

8. REFERENCES

[1] J. L. Caldwell, "Implementation of short-time homomorphic dereverberation," Master's thesis, MIT, 1971.

[2] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity based blind dereverberation for single channel speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 80–95, 2007.

[3] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[4] A. Krüger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," in *Interspeech*, 2009, pp. 1231–1234.

[5] H. Kameoka, T. Nakatani, and T. Yohioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *ICASSP*, 2009, pp. 45–48.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *NIPS workshop on advances in models for acoustic processing*, 2006.

[7] S. M. Gabriel and M. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," in *ICASSP*, 2001.

[8] M. Shashanka, "Latent variable framework for modeling and separating single channel acoustic source," Ph.D. dissertation, Boston University, 2007.

[9] R. Ratnam, D. L. Jones, B. C. Wheeler, and W. D. O'Brien, "Blind estimation of reverberation time," *Journal of the Acoustic Society of America*, vol. 114, 2003.