Blind Dereverberation of Audio Signals

E4810 Final Project

Graham Grindlay
grindlay@ee.columbia.edu

December 27, 2008

Abstract

This project examines the problem of single channel blind dereverberation. After estimating the T60 value, a time-domain binary masking approach was used to remove regions of the signal that were largely dominated by reverberant energy. Performance of the system was examined for several different classes of audio (hand clapping, drums, and speech) and for varying amounts of reverberation. In addition to subjective evaluation via spectrograms and audio output, results were assessed by comparing the accuracy of the binary mask to an ideal binary mask calculated using a clean reference signal. Results were generally good although as expected, performance was better for sparse signals. Interestingly, some mask accuracy results were better when more reverberation was present, even though subjectively the system performed worse on these examples.

1 Introduction

1.1 Background

Although reverberation can be desirable in some situations, there are many practical applications where it causes significant problems. For example, it is well known that the performance of automatic speech recognition systems suffers greatly when reverb is present and it is likely that similar problems arise for automatic music transcription. Reverberation can also impact speech intelligibility, especially for hearing-impaired listeners [4].

In this project, I explore a blind approach to dereverberation of audio, meaning that no prior knowledge about the clean signal or transfer function is assumed. Although there are other approaches to this difficult problem [9, 8], many require multiple audio channels [3] or make assumptions about the harmonicity of the signals [5]. In contrast, the approach taken here works with single channel recordings, makes no assumptions about the source, and is conceptually straightforward.

1.2 Motivation

My approach to blind dereverberation is motivated by the simple observation that the envelope of late reflections in a reverberant signal is well described by

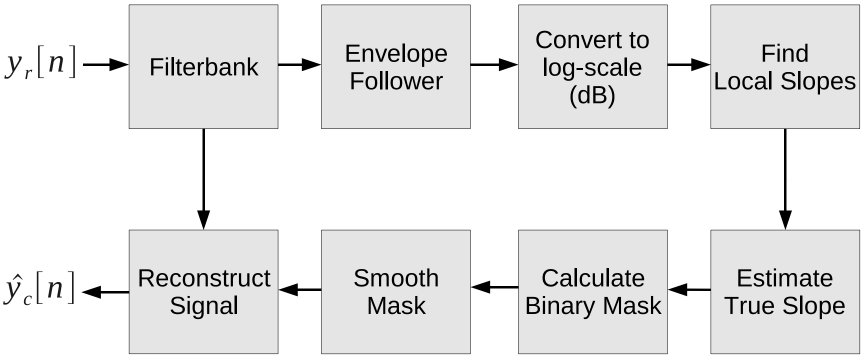

Figure 1: Overview of the proposed dereverberation algorithm

an exponential decay curve: $e_r[n] \approx \exp(-1/\tau)^n = a^n$. Therefore a describes the reverberation time of a room or environment and relates to the T60 time (the time it takes an impulse response to drop 60dB below its initial level) as $\text{T60} = -\frac{\log(1000)}{\log(a)}$. If we can estimate a, we can estimate the T60 time which can be used to identify and remove regions of the signal that are mostly reverberant energy.

2 Algorithm

This section describes the proposed algorithm which is depicted in Figure 1. The very first step is to run the reverberant signal through a bank of bandpass filters. Because reverberation times are frequency dependent, the algorithm is run on each band individually and the results are combined at the end. Next the envelope of each band is calculated and transformed to the dB scale. A window of length N (empirically determined) is then moved along the envelope signal one sample at a time and at each position the window of samples is fit using linear regression. The slope of the line fit for each window is retained and the set of all slopes is postprocessed to determined the estimate of the true slope (which in the linear domain corresponds to a scaled version of the exponential decay parameter, a).

2.1 Filterbank

Several different types of filterbank were implemented in order to investigate the impact on dereverberation performance. This included both linearly-spaced and constant-Q 4^{th}-order Butterworth bandpass filters, constant-Q linear phase FIR filters (designed using Parks-McClellan), and Gammatone filters (designed using Glasberg and Moore parameters [2] with Nick Clark's Gammatone toolbox [1]). For the Butterworth and Gammatone filterbanks, phase distortion was minimized using time-reversed filtering (via Matlab's *filtfilt* function) while for the FIR filterbank, shifting was used to correct for the introduced delay.

In general, I found that the constant-Q Butterworth filters produced the best results for speech, while linearly spaced Butterworth filters worked better for drum and clapping sounds. It is important to note, however, that this assessment is based largely on the binary mask accuracy metric described in Section 2.5. In terms of perceived quality, the filter types were similar. Gammatone filters produced results that were very similar to the linearly spaced Butterworth filterbank.

2.2 Envelope Follower

Two approaches to envelope detection were implemented: a simple low-pass filter and rectification scheme as well as a detector based on the Hilbert transform. In general, the latter approach seemed to work better and so it was used in the experiments presented below. Using $\hat{y}[n]$ to denote the Hilbert transform of $y[n]$, the Hilbert transform based envelope detector is defined as:

$$e[n] = \sqrt{\hat{y}[n]\ \hat{y}^*[n]}$$

The result of the detector was then low-pass filtered using a first-order Butterworth filter with a (empirically determined) cutoff frequency of 80Hz. This helped to remove high frequency elements of the envelope that were not related to reverberation. The time reversal zero-phase filtering scheme was used to eliminate the phase distortions that the lowpass filter would have otherwise introduced.

2.3 Analysis

The first analysis step is to convert the envelope to a logarithmic (dB) scale. Next a window of length N is moved along the envelope and linear regression is used to fit a line at each position. The slopes $s_i \in \mathbf{s}$ of the lines fit at each window position are then analyzed to determine the best estimate of the true slope, s^*. Following Ratnum et al. [6], this was done using an order-statistics approach where a threshold value of s was selected such that the left tail of the density function $p(s)$ occupies a specified percentile value, γ. Although this method of selecting the estimated value is somewhat ad-hoc, it works reasonably well in practice. I found that a value of $\gamma = 0.25$ worked well in most circumstances.

2.4 Masking

Regions of the input signal that are largely reverberant decay are removed using a masking technique. First a binary mask is constructed from the set of local slope calculations. Using α to represent an empirically determined tolerance parameter, mask cell i is defined as:

$$m_i = \begin{cases} 0 & \text{if } |s_i - s^*| < \alpha \\ 1 & \text{otherwise} \end{cases}$$

In practice I found that binary masking was too abrupt and introduced unacceptable transient artifacts. To help alleviate this problem, a moving average filter (50ms worked well in practice) was used to smooth the mask.

2.5 Evaluation

As mentioned earlier, along with spectrograms and listening tests, an objective metric based on binary mask accuracy was used to evaluate the results of the dereverberation algorithm. The calculated mask m was compared to the ideal binary mask $\hat{m}$ which was calculated using the clean and reverberant signal envelopes, e_c and e_r. Using β to denote an empirically set threshold parameter, $\hat{m}$ is defined as:

$$\hat{m}_i = \begin{cases} 0 & \text{if } \frac{e_r}{e_c} > \beta \\ 1 & \text{otherwise} \end{cases}$$

Evaluating the accuracy of a mask m is then simply defined to be the percentage of points in m that match $\hat{m}$.

3 Experiments

Table 3 summarizes the experimental results. All spectrograms, binary mask figures, audio examples, and source code will be available at:

`http://www.ee.columbia.edu/~grindlay/classes/E4810/project.html`

Table 1: Summary of accuracies obtained for different sound classes with different levels of reverberation

	T60=0.25sec	T60=0.5sec	T60=1.0sec
Clapping	0.71	0.86	0.97
Drums	0.82	0.86	0.87
Digits(F)	0.63	0.71	0.78
Digits(M)	0.68	0.73	0.78
Speech(F)	0.73	0.73	0.74
Speech(M)	0.72	0.74	0.73

3.1 Reverb IRs

In order to precisely control the level of reverberation present in the test signals, artificial impulse responses were used. The impulse responses were constructed by shaping a random series of filter taps (i.e. -1, 0, or 1) with an exponential curve. Typically, the probability of non-zero taps was $1/10^{th}$ of the zero tap density. The exponential curve parameter τ was determined by the desired T60 value.

I also experimented with some recorded impulse responses as well as impulse responses generated by commercial reverberators. In these experiments, Schroeder's method [7] was used to estimate the T60 of the impulse response. The results were generally similar to those when using artificial impulse responses of the same T60 time and so are not included in the results below.

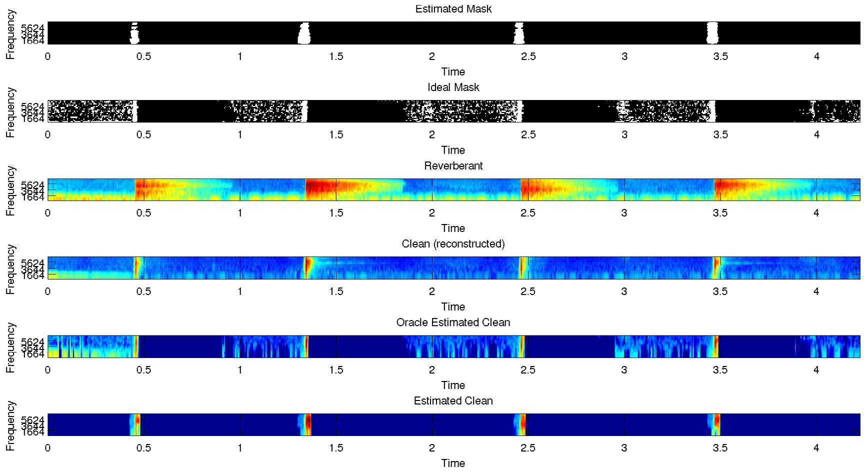

Figure 2: Clapping results with T60 = 0.5 seconds

3.2 Clapping

Hand claps provide a good first test as they closely approximate impulse responses and should therefore be easy to dereverberate. Artificial reverb IRs were constructed as described in Section 3.1 and added to an anechoic recording of four hand claps using convolution. Linear Butterworth filters were used in the filterbank, the analysis window length was set to 100ms, α to 0.05, and β to 2. Because of the impulse response characteristics of the clapping sounds, the binary mask did not require smoothing.

Mask accuracy results for three different T60 values are given in Table 3. Interestingly, in terms of mask accuracy, the algorithm performed best when the most reverb was present, a somewhat counterintuitive finding. However, much of this difference appears to be due to silent portions of the sound clip and the fact that the ideal mask is noisy in these areas. Figure 2 shows the results for T60 = 0.5 seconds which are, from top to bottom: calculated binary mask, ideal binary mask, the reverberant sound, the original clean sound, the reverberant sound when cleaned with the ideal binary mask, and the reverberant sound when cleaned with the algorithm's binary mask.

3.3 Drums

A slightly more challenging drum clip was also tested with varying amounts of reverberation. The clip has bass drum, high-hat, and snare drum parts present which overlap. The filterbank used linear Butterworth filters, the analysis window was 100ms, α was set to 0.015, and β set to 2. Again, we see the same trend of higher mask accuracy with higher T60 values, even though these results sound markedly worse than those of the lower T60 values.

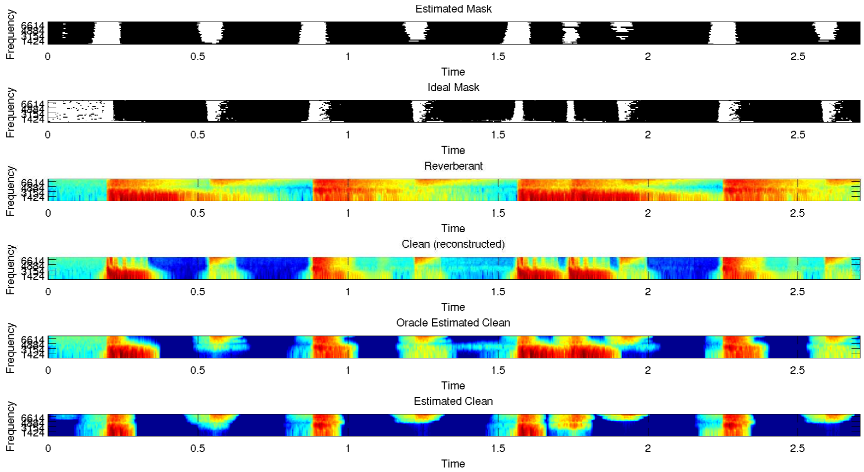

Figure 3: Drum results with T60 = 0.5 seconds

3.4 Speech

3.4.1 Digits

Percussive sounds are easy to dereverberate becausethey are close to impulse responses. Speech signals, on the other hand, are more difficult to dereverberate because they are often dense. I constructed sound clips of both male and female speakers saying the digits 'one' through 'four' in succession. The digits were placed far enough apart that there was time for reverberant decay between them. This made the digit task challenging, but not as difficult as a naturally uttered sentence.

Again we find the same counterintuitive relationship between T60 and mask accuracy as was found for drums and clapping. This time, however, the results do not sound appreciably different (although T60 = 0.25 seconds might be a bit better than the other values). Figures 4 and 5 show the results for male and female speakers using constant-Q Butterworth filters in the filterbank and the same parameter settings as the drum experiment.

3.4.2 Sentences

The most challenging dereverberation experiment used natural (in so far as the TIMIT sentences are natural!) speech utterances. The sentences, "she had your dark suit in greasy wash water all year" (female speaker) and "cottage cheese and chives is delicious" (male speaker) were used. Numerical results are given in Table 3 and the spectrograms and masks for T60 = 0.5 seconds are shown in Figures 6 and 7. Interestingly, in this case the mask accuracy performance is relatively consistent across T60 values. Listening to the results, however, it is clear that the speech becomes more distorted as the reverberation level increases, although the degree to which the reverb has been attenuated also becomes more apparent with the more reverberated examples.

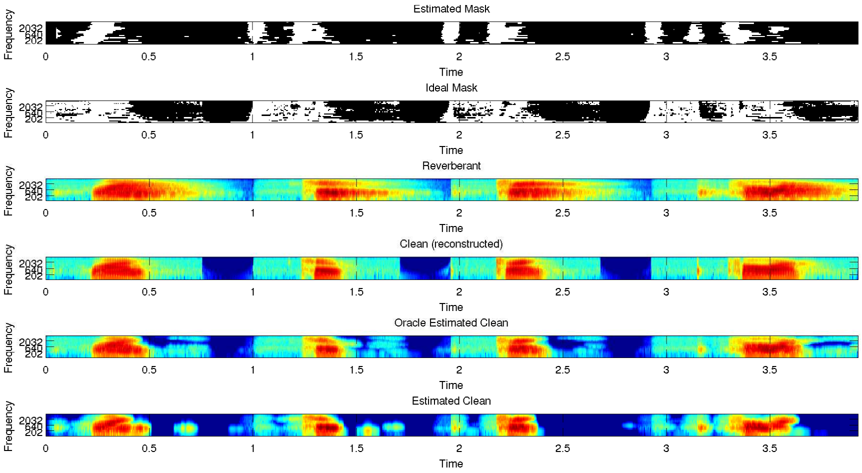

Figure 4: Female digit results with T60 = 0.5 seconds

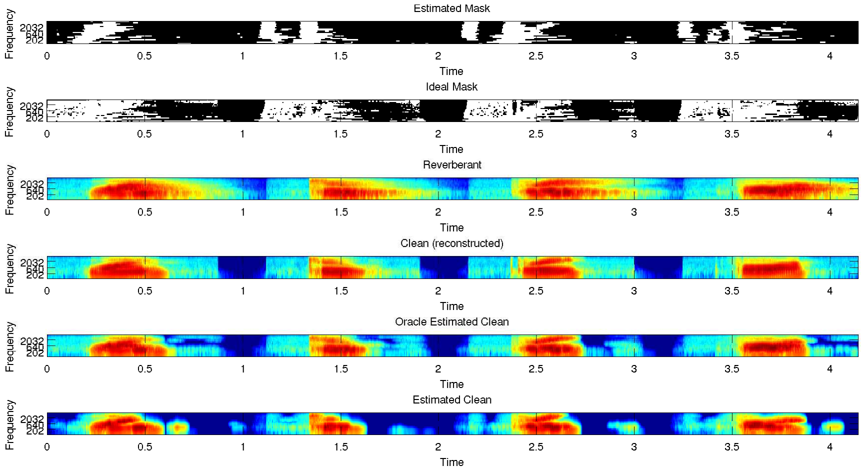

Figure 5: Male digit results with T60 = 0.5 seconds

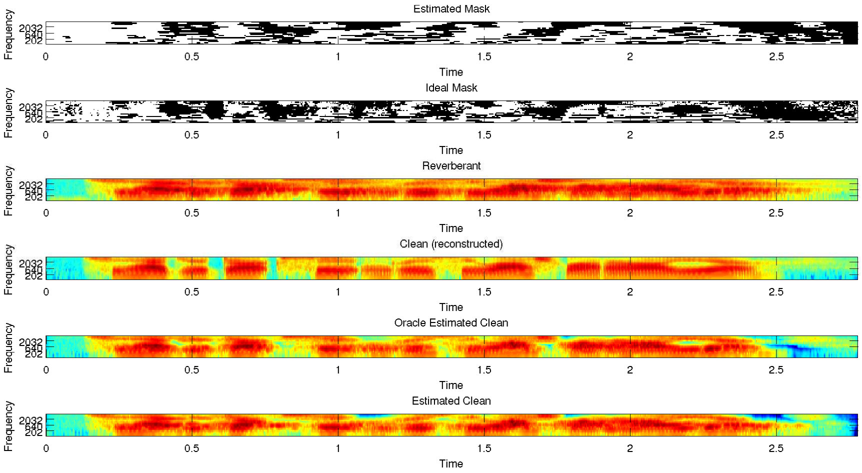

Figure 6: Female sentence results with T60 = 0.5 seconds

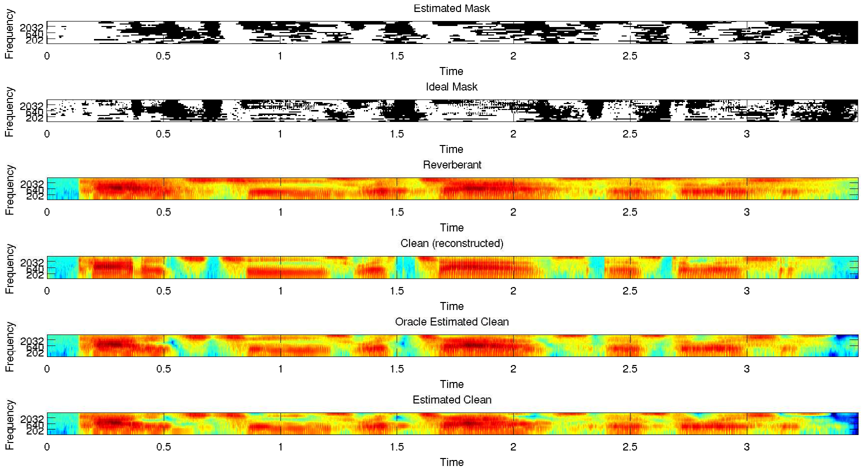

Figure 7: Male sentence results with T60 = 0.5 seconds

4 Conclusions

This simple approach to dereverberation seems to have worked reasonably well. Although the algorithm distorted some elements of the signals it operated on, it did manage to significantly reduce reverberant energy in many cases. Not surprisingly, more percussive and impulse response-like signals, such as clapping and drumming, tend to work better than dense speech.

It seems likely that the system could be improved in several areas. First, the quantitative mask accuracy metric does not always yield results that correspond to subjective assessment. Although this is, to some extent, the point of an objective metric, there are cases where the mask accuracy results are clearly off. Second, although not much time was spent tuning the algorithm's parameters, performance can be quite sensitive to thir settings, especially α. It would be helpful to have a data-driven approach to determining these parameters. Finally, although it would complicate the algorithm, it may be possible to include some kind of source modeling. This would allow prior knowledge to be incorporated and could reduce the harmonic artifacts that sometimes result from the masking process.

References

[1] N. Clark. Gammatone toolkit v2.0. `http://www.mathworks.com/matlabcentral/fileexchange/15313`, 2008.

[2] B. Glasberg and B. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, 1990.

[3] M. Miyoshi and Y. Kaneda. Inverse filtering of room impulse response. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):145–152, 1988.

[4] A. Nabelek. *Acoustical Factors Affecting Hearing Air Performance*, chapter Communication in Noisy and Reverberant Environments. Allyn and Bacon, 1993.

[5] T. Nakatani and M. Miyoshi. Blind derverberation of single channel speech signal based on harmonic structure. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 92–95, 2003.

[6] R. Ratnam, D. Jones, C. Wheeler, and D. O'Brien. Blind estimation of reverberation time. *Journal of the Acoustical Society of America*, 114(5):2877–2892, 2003.

[7] M. Schroeder. New method of measuring reverberation time. *Journal of the Acoustical Society of America*, 37:409–412, 1965.

[8] M. Wu and D. Wang. A one microphone algorithm for reverberant speech enhancement. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 884–887, 2003.

[9] M. Wu and D. Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):774–784, 2006.