Audio Engineering Society

Convention Paper 9905

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Blind Estimation of the Reverberation Fingerprint of Unknown Acoustic Environments

Prateek Murgai[1], Mark Rau[1], and Jean-Marc Jot[2]

[1]*Center for Computer Research in Music and Acoustics (CCRMA), Stanford University*
[2]*Magic Leap*

Correspondence should be addressed to Prateek Murgai (pmurgai@ccrma.stanford.edu)

ABSTRACT

Methods for blind estimation of a room's reverberation properties have been proposed for applications including speech dereverberation and audio forensics. In this paper, we evaluate algorithms for online estimation of a room's "reverberation fingerprint", defined by its volume and its frequency-dependent diffuse reverberation decay time. Both quantities are derived adaptively by analyzing a single-microphone reverberant signal recording, without access to acoustic source reference signals. The accuracy and convergence of the proposed techniques is evaluated experimentally against the ground truth obtained from geometric and impulse response data. The motivations of the present study include the development of improved headphone 3D audio rendering techniques for mobile computing devices.

1 Introduction

Blind reverberation identification has been the object of extensive research targeting applications to speech communication and speech recognition [1]. More recently, it was studied in the context of audio forensics [2] and suggested for controlling headphone 3D audio artificial reverberation processing for augmented reality applications [3].

In these applications, it is desirable to estimate certain reverberation properties of a local environment by analyzing a single-microphone recording of the reverberant signal, although no reference/source signal is available to the system. The definition of the "roomprint" [2] or "reverberation fingerprint" [3] includes the room reverberation time vs. frequency (i.e. measured in sub-bands) and the reverberation power level (or, alternatively, the room's cubic volume). An advantage of such a characterization is that, in typical enclosures, the reverberation time is independent of source or receiver position, directivity or orientation, and thus may be considered a property of the room itself. On the other hand, these source and receiver parameters may affect reverberation power and spectrum [3] [4], and must be eliminated or compensated for when it is desired to characterize the room itself. In [3], it is proposed to isolate the effect of the room by restricting the "reverberation fingerprint" characterization to the diffuse part of the reverberation decay and exploiting a stochastic reverberation model previously proposed in [4].

Augmented (or Mixed) reality (AR/MR) presents a new challenge where the space in which the user is situated may have audibly different acoustic properties than the space inhabited by the virtual audio sources. Jot and Lee proposed a method to alter a known room response so as to simulate a room having a different reverberation fingerprint, in order to achieve a more natural virtual 3D audio reproduction [3]. This method assumes that the local acoustic environment's reverberation fingerprint is known but does not propose an online technique for measuring it. The present work focuses on the blind online estimation of reverberation decay time and room volume, applicable to augmented reality audio applications.

The sub-band reverberation time is an essential measure for characterizing the acoustics of a space and quantifying the subjective persistence of a sound source within that space, thus making its estimation of prime importance for AR applications. As explicit measurement of reverberation decay time using test signals is not practical while using AR applications in unknown spaces, we propose a method which blindly estimates the reverberation time by passively listening to the sound sources that are already present in the space. Several previous studies describe offline methods for the blind estimation of the sub-band reverberation time from recorded speech signals [5] [6] [7]. We build on this prior work to propose an algorithm that employs a running energy envelope estimator coupled with a peak detection algorithm in order to segment intermittent decays that occur in a continuous speech signal recording. The method assumes that the reverberation time will be revealed by the fastest decay observed among the detected decay segments, and converges to a reverberation time estimate based on that assumption. In the following, an online implementation of this decay time estimation technique is described.

The diffuse-field reverberation power can be assumed to be independent of source or listener positions in a "well-behaved" (or "mixing") room. For a given decay time, source and listener, the power of the diffuse field reverberation is inversely proportional to the cubic volume of the room [4]. If the volume and reverberation time of an unknown room are detected, a reverberation impulse response measured in a known room can be modified and scaled accordingly, in order to match the unknown room's diffuse field reverberation power, as proposed in [3]. Statistical methods similar to that of Shabtai [8] may be applicable to the blind prediction of the cubic volume of an unknown room from locally recorded speech signals. In the following, we explore an adaptation of this approach to online blind prediction of diffuse-field reverberation power.

2 Methods

2.1 Reverberation Decay Rate Prediction

A blind reverberation time estimation technique that monitors intermittent decays in running speech is proposed. The method assumes that, in a given frequency band, the reverberation time is derived from the fastest decay rate observed among all the energy envelope decays that occur in the recorded speech signal. The technique is broken down into the following steps.

2.1.1 Energy Envelope Estimation

In a speech signal recorded in a room, the rate of envelope decays is limited (slowed down) by the room's reverberation decay rate. The reverberation time can be measured during speech pauses and determines the fastest decay rate observed during signal energy envelope decay segments. In order to detect each of these decay segments, we first employ a running energy estimator based on Equation 1.

$$e[k] = \alpha \cdot x[k-1]^2 + (1-\alpha) \cdot x[k]^2 \qquad (1)$$

Where x is the input speech signal, $e[k]$ is the estimated running energy and α is the forgetting factor which estimates the weighting between the current and past sample in power calculation. In our experiment, α is set to 0.2.

Once the running energy has been estimated, a leaky integrator based root mean square detector is used to approximate its envelope. The envelope detector is given by Equation 2.

$$y[k] = \sqrt[\frac{1}{2}]{\beta \cdot y[k-1]^2 + (1-\beta) \cdot e[k]^2} \qquad (2)$$

Where $y[k]$ is the current sample of the detector and $e[k]$ is the current sample of the running energy derived from Equation 1. $\beta = e^{\frac{-1}{\tau \cdot Fs}}$, Fs is the sampling rate for the input speech signal and τ is the time constant for the detector. To ensure that the detector follows the input closely, τ is assigned a small value of 5 milliseconds.

Fig. 1: Blind reverberation time estimation process

2.1.2 Decay Start and Stop Estimation

Our next task is to locate the points in the energy envelope where each decay begin occurring. For this purpose, we design a simple local peak estimation technique which looks at the immediate past and future neighboring $e[k]$ samples to designate the current sample in the energy envelope as a local peak.

We further select the peaks based on a threshold amplitude value that depends on the largest peak previously detected in the energy envelope. Once the decay start points have been estimated, we designate each peak as a valid start point only if the signal beyond this point decreases monotonically after a duration at least equal to a threshold amount of time given by $\tau_{threshold}$. For our application, we set $\tau_{threshold}$ to 50 milliseconds. The times where the signal stops to monotonically decrease are tagged as decay stop points.

2.1.3 Intermittent Decay Time Estimation

We fit a straight line characterized by slope γ to each of the legitimate decays in the running energy envelope (on a dB scale). Then the decay time t_d for a given decaying region is approximately given by:

$$t_d \approx \frac{-60}{\gamma} \tag{3}$$

Equation 3 estimates the time it takes for the detected signal to drop by 60 dB. As a 60-dB drop is usually not available, we use Equation 3 to find the approximate decay time.

Depending on the number of legitimate decays detected, we build a set which stores all the decay times that have been observed in the running signal:

$$T = \{t_{d1}, t_{d2}, t_{d3}, t_{d4} \dots\} \tag{4}$$

Based on our assumption that the reverberation time (t_{RT}) is the fastest decay among all the decays, t_{RT} will be given by:

$$t_{RT} = min(T) \tag{5}$$

Figure 1 depicts the complete blind reverberation time estimation process.

2.1.4 Reverberation Time Selection From Minimum Decay Times

While we are deriving minimum decay times from a running speech signal, it is important that, before designating a decay time as the reverberation time of the acoustic space, we obtain a sufficient recurrence of consistent decay rate estimates. For this purpose, we save all the minimum decay times and if the number of minimum decay times which fall within a tolerance limit ε is greater than η then the final reverberation time is given by the average of those decay times. In our experiment, ε is set to 0.1 and η is set to 5. These values are tuned based on multiple experimental runs. The results below illustrate the speed of convergence observed with this technique.

2.2 Room Volume Prediction

A statistical approach similar to that of Shabtai et. al. [8] was taken to predict the cubic volume of an unknown room from speech recordings. Acoustic features related to the volume of a room were calculated based on impulse response approximations from reverberant speech in known rooms. These features were used to train a predictive model.

2.2.1 Feature Extraction

Room impulse response measurements captured in a quiet room using reliable equipment would be ideal, but not practical in a real world setting. Here, a substitute for the room impulse response is taken as the signal following a speech utterance. While the speech signal does not resemble an impulse, this subtitution will still provide some insight into the room's impulse response characteristics. To find the temporal locations where speech fragments are stopped, a root mean square level detector was fit to the signal as described in Section 2.1.1. When no increases in level are detected over a certain time threshold, a “speech stop" was deemed found. An example of this process on sample speech audio is shown in Figure 2.

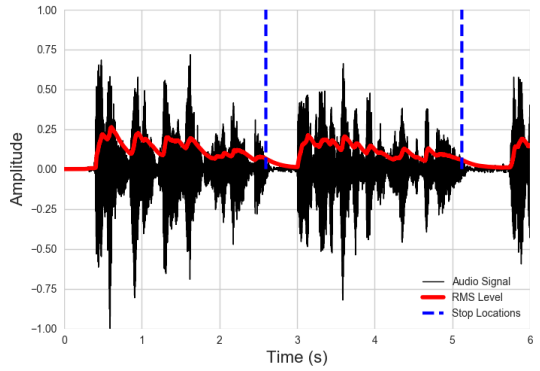

Fig. 2: Example of energy envelope based speech stop detection.

The acoustic features used were T_{60}, C_{80}, C_{50}, density of early reflections, kurtosis in the time domain, density of low frequency room modes, and kurtosis in the frequency domain. These features were chosen as their definitions are related to the room volume [9]. Frequency-dependent T_{60}, C_{80}, and C_{50} measures were calculated in octave bands from 20 Hz to 20 kHz.

The density of early reflections is related to the room volume, as there will be a higher number of early reflections in a small room compared to large room. An approximation to the density of early reflections was made by calculating the number of local peaks in the first 100 ms of the impulse response. A high kurtosis value means that more of the signal variance is due to infrequent, large signal variations, while a low value means that there are frequent small signal fluctuations, hence the kurtosis is related to the room volume. The time domain kurtosis was calculated over the first 100 ms of the estimated impulse response.

The density of low frequency modes is related to the room volume, as a small room will have room modes starting at higher frequencies. The density of low frequency room modes was estimated by calculating the number of local peaks in the lower 20% of the frequency spectrum. The frequency-domain kurtosis was also measured in the lower 20% of the frequency spectrum.

2.2.2 Prediction

A Gaussian process regression model was used to predict the volume of an unknown room from speech data. Gaussian process is a supervised learning method which uses a measure of the similarity between points to predict the value of an unknown point from training data [10]. The extracted features and volumes of known rooms were used to train the model. The same features are extracted from a speech recording made in an unknown room, and the classifier is used to predict the volume of this room. Since real speech is continuous, multiple volume predictions can be made as time progresses. To achieve a more reliable volume prediction, a running average of volume predictions is used as a more reliable volume prediction metric.

3 Experimental Setup and Results

3.1 Reverberation Decay Rate Prediction

To perform our experiments, we employ impulse responses from the Open AIR Library[1] [11], convolve them with anechoic speech recordings and use the convolved outputs to blindly estimate the reverberation times. As reverberation times are frequency dependent, we filter the convolved outputs with an octave filter and run our blind estimation technique on each of the filtered output signals. Figure 3 shows the comparison between the actual reverberation times and the blind estimates for four different acoustic spaces.

Next, we will look at how fast the blind estimate converges towards stable reverberation values, as this would be an important factor while designing real time applications.

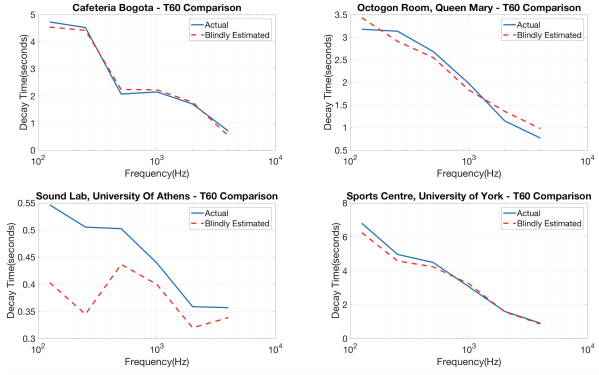

Fig. 3: Comparison between blind estimate and actual reverberation time

[1] openairlib.net

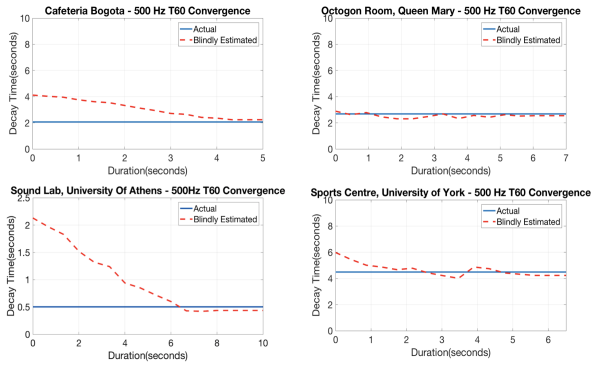

Fig. 4: Convergence of reverberation time estimate (500Hz frequency band)

Figure 4 illustrates the speed of convergence of the reverberation time estimate, in the four acoustic spaces considered in Figure 3.

3.2 Room Volume Prediction

The volume prediction method was tested with synthesized room impulse responses generated with the image source method using a MATLAB toolbox [12]. A total of 154 shoe box shaped rooms ranging in volume from 8 m^3 to 932 m^3 were synthesized. The impulse responses were convolved with 12 recordings of anechoic speech, each 15 seconds in length. This resulted in 1848 total speech samples. The speech samples were then analyzed to find locations where the speech was stopped, resulting in an impulse response approximation. A total of 4380 room impulse response approximations were found, and the features mentioned in Section 2.2 were extracted. These features and the associated room volumes were used to train a Gaussian process regression model which was implemented in Python using the scikit-learn toolbox [13].

To test the prediction model, a set of 51 room impulse responses over the same range, but having different volumes was generated. As with the training, these room impulse responses were convolved with anechoic speech data, speech stops were found, and the features were extracted. Volume predictions were made and then averages were taken for each speech signal used. Figure 5 shows volume predictions and the true labels over a range of volumes for speech signals of length 15 seconds.

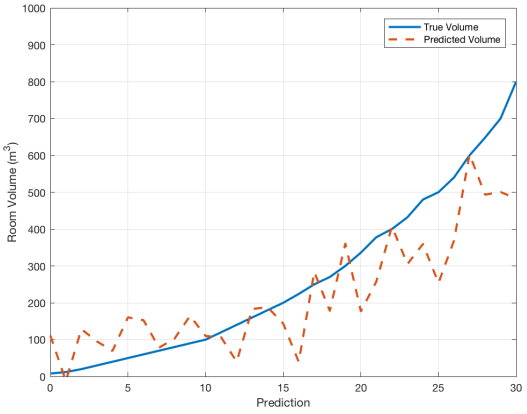

Fig. 5: Room volume estimates compared to actual volumes, for various room sizes.

Figure 6 illustrates how the volume prediction evolves over time as more predictions are made. The top and bottom graphs show the average estimated volumes for 200 m^3 and 400 m^3 rooms, respectively. There were 56 volume predictions obtained over a 3.5 minute reverberant speech recording.

4 Summary and Discussion

We present a preliminary study exploring the feasibility of continuous blind estimation of the reverberation fingerprint of an unknown room by monitoring recorded speech signals. To this aim, we developed online variants of previously published methods for the blind estimation of a room's cubic volume and of its sub-band reverberation time. We tested the proposed methods with recorded speech samples convolved with measured or calculated room impulse responses.

For reverberation time estimation, our initial results are encouraging, both in terms of convergence time and in terms of accuracy, especially in rooms having relatively long reverberation times. Our results suggest that an improvement of the proposed online blind estimation method would be beneficial in order to avoid underestimation bias for rooms having shorter reverberation times, and improve accuracy at low frequencies.

Generalizing tentatively from the presented results, it seems realizable to achieve convergence to a close estimate of the actual reverberation time after just a few seconds of speech activity in the room.

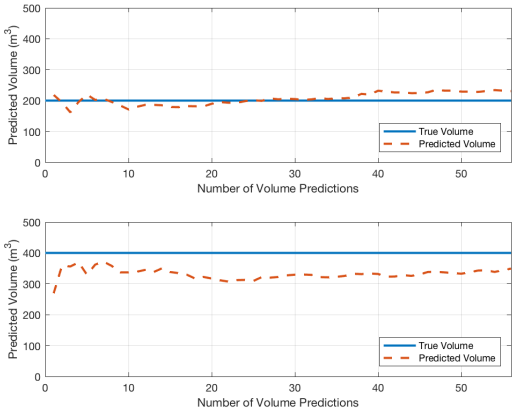

Fig. 6: Room volume estimate with increasing number of predictions, in two rooms having volumes 200 m^3 (top) and 400 m^3 (bottom).

Room volume estimation based solely on the analysis of speech recordings is a challenging task. The method explored here, based on a machine learning approach, nevertheless tentatively appears capable of converging, after 10 seconds of speech activity in the room, to within 25 percent of the actual volume (which translates to approximately a 1 dB error in reverberation power). Further experiments will be necessary in order to verify whether this objective is attainable in typical indoor environments.

References

[1] Naylor, P. and Gaubitch, N., *Speech Dereverberation*, Berlin, Germany: Springer-Verlag, 2010.

[2] Moore, A. H., Brookes, M., and Naylor, P. A., "Room Identification Using Roomprints," in *Proc. Audio Engineering Society 54th International Conference: Audio Forensics*, 2014.

[3] Jot, J.-M. and Lee, K. S., "Augmented Reality Headphone Environment Rendering," in *Proc. Audio Engineering Society Conference on Audio for Virtual and Augmented Reality*, 2016.

[4] Jot, J.-M., Cerveau, L., and Warusfel, O., "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Proc. 103rd Audio Engineering Society Convention*, 1997.

[5] Prego, T. d. M., de Lima, A. A., Netto, S. L., Lee, B., Said, A., Schafer, R. W., and Kalker, T., "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *Journal of the Acoustical Society of America*, 131(4), pp. 2811–2816, 2012.

[6] Eaton, J., Gaubitch, N. D., and Naylor, P. A., "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, 2013.

[7] Diether, S., Bruderer, L., Streich, A., and Loeliger, H.-A., "Efficient blind estimation of subband reverberation time from speech in non-diffuse environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 743–747, 2015.

[8] Shabtai, N., Rafaely, B., and Zigel, Y., "Room volume classification from reverberant speech," in *Proc. International Workshop on Acoustic Signal Enhancement*, 2010.

[9] Shabtai, N. R., Zigel, Y., and Rafaely, B., "Room volume classification from room impulse response using statistical pattern recognition and feature selection," *Journal of the Acoustical Society of America*, 128(3), pp. 1155–1162, 2010.

[10] Rasmussen, C. E. and Williams, C. K., *Gaussian processes for machine learning*, volume 1, MIT Press, 2006.

[11] Murphy, D. T. and Shelley, S., "Openair: An interactive auralization web resource and database," in *Proc. 129th Audio Engineering Society Convention*, 2010.

[12] Wabnitz, A., Epain, N., Jin, C., and Van Schaik, A., "Room acoustics simulation for multichannel microphone arrays," in *Proc. International Symposium on Room Acoustics*, pp. 1–6, 2010.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 12(Oct), pp. 2825–2830, 2011.