Real-World Evaluation of Multichannel Audio Enhancement Using Acoustic Pilot Signals

Ryan M. Corey and Andrew C. Singer
University of Illinois at Urbana-Champaign

Abstract—In reverberant environments, the performance of multichannel audio enhancement systems depends on the accuracy with which the acoustic impulse responses between sources and microphones can be estimated. Since these responses are difficult to estimate in real-world environments, many multichannel audio enhancement experiments are conducted using simulated data. We propose an experimental method that sidesteps the channel estimation problem by measuring the acoustic impulse responses using pilot signals emitted from a beacon placed near the source of interest. We demonstrate the utility of the method on a multichannel speech mixture recorded in a noisy and reverberant room.

Index Terms—Audio enhancement, microphone array processing, beamforming, source separation, system identification

I. INTRODUCTION

Multichannel audio enhancement technology can improve the ability of both humans and machines to listen to sounds in challenging noisy environments. Microphone arrays embedded in mobile phones, teleconferencing equipment, cars, home electronics, and wearable devices such as hearing aids enable an audio processing system to distinguish between sounds arriving from different directions and enhance the received signal in a variety of ways [1]–[3]. Here, we focus on one common scenario: an array is used to isolate a single speech source from a noisy mixture, attenuating all other sound sources and any internal system noise as much as possible.

A typical multichannel audio enhancement system is shown in Figure 1. The sounds produced by each source propagate through the room to an array of sensors. Each source-microphone pair is associated with a unique acoustic path that depends on their spatial locations. An estimation algorithm estimates some parameters of this acoustic path. Those parameters are used to design a spatial filter to isolate the source of interest and attenuate noise. Multichannel audio enhancement methods differ in the parameters they use to model the channel and in the criteria they use to design the spatial filter. The simplest methods are based on a far-field anechoic model, in which the source-to-microphone channels are pure time delays that depend only on the direction of arrival (DOA) of the source. This model is often paired with a delay-and-sum multichannel filter, which is also known as a beamformer since its gain pattern forms a beam toward the target source [4].

However, most audio enhancement systems are used indoors, where sources and microphones are closely spaced and there is significant reverberation. In a reverberant room, source signals arrive at the array from multiple directions, so it no longer makes sense to point a beam in a single direction. Instead, we must design multichannel filters that apply different gains to signals with different reflection patterns, known as acoustic impulse responses (AIR) [5]. These filters can be more powerful and versatile than simple beamformers; for example, the increased spatial diversity provided by reverberation allows the array to separate sources from similar directions [6], [7]. Unfortunately, since typical room reverberation times are hundreds of milliseconds long, it is difficult to estimate the AIRs of a source from mixture data alone [8].

Indeed, parameter estimation is considered a limiting factor in multichannel reverberant audio enhancement [9]. There are a number of interesting research questions about multichannel filter design that cannot be explored experimentally because of the parameter estimation bottleneck: how does the performance of large microphone arrays scale in real rooms? How should we select filter coefficients to trade off noise reduction performance, mismatch sensitivity, computational complexity, and processing artifacts? Given better parameter estimation algorithms, how well could multichannel audio enhancement perform and what new applications would it enable?

To evaluate the performance of multichannel filters and other audio enhancement techniques, researchers have typi-
cally relied on one of three experimental methods, each with its own strengths and weaknesses:

1) Fully synthetic mixtures: The AIRs are measured and stored, then convolved with anechoic source recordings to generate an artificial mixture. A number of impulse response datasets are publicly available for this purpose [10], [11]. This method allows researchers to precisely control the acoustic conditions and track the contributions of all the source and noise signals to the filter output and to compare the performance of the algorithms under study to ideal filters designed using ground-truth information [12], [13]. However, these synthetic mixtures are the least realistic of the three methods and the results may not generalize to real-world systems.

2) Separately recorded sources: Several sound sources are generated and recorded individually in the same room. The recordings are then added together to form an artificial mixture. Because the sources are recorded in a real room, these mixtures are more realistic than fully synthetic mixtures, and are often used in source separation challenges, e.g. [14]. However, because each source must be recorded in complete isolation, such data can only be collected in carefully controlled laboratory conditions, not in the real-world environments where audio enhancement systems are most urgently needed.

3) Simultaneously recorded sources: Audio data is recorded in a real-world environment with all sound sources active at once. This data is the most realistic and can validate the real-world performance of an enhancement algorithm. It is also easy to collect and many datasets are available, e.g. [15]. However, since there is no ground truth data, it is impossible to evaluate algorithm performance objectively. Furthermore, researchers cannot compare the output to that of an ideal filter designed using the true channel parameters.

In this work, we propose an experimental method to evaluate multichannel audio enhancement in real-world environments while sidestepping the difficult parameter estimation problem. The method is inspired by wireless communication systems, which nearly always use a pilot signal to measure the channel before transmitting data. An acoustic transmitter placed near the sound source—typically a human talker—emits a known pilot signal that is used to measure the channel, as shown in Figure 2. This measured channel is then used to design multichannel filters. This method allows us to design more selective filters than we could otherwise in reverberant environments and demonstrates the achievable performance of multichannel audio enhancement systems.

II. AUDIO ENHANCEMENT WITH PILOT SIGNALS

A. Multichannel audio enhancement

Consider an array of $M$ microphones. Let $\mathbf{x}[n] = [x_1[n], x_2[n], \ldots, x_M[n]]^T$ be the vector signal recorded from the array, where $n$ is the discrete-time index. We assume a convolutional mixing model:

$$\mathbf{x}[n] = \sum_k \mathbf{a}[k]s[n-k] + \mathbf{z}[n], \quad (1)$$

where $s[n]$ is the desired source signal, $\mathbf{a}[n]$ is the vector of discrete-time AIRs from the source to the array, and $\mathbf{z}[n]$ is the noise vector due to unwanted sources. The output signal $y[n]$ is produced by multichannel filtering,

$$y[n] = \sum_k \mathbf{w}^T[k]\mathbf{x}[n-k], \quad (2)$$

where $\mathbf{w}[n]$ is the vector sequence of filter coefficients. These coefficients are typically the solution to an optimization problem that depends on $\mathbf{a}[n]$ and the statistics of $s[n]$ and $\mathbf{z}[n]$.

Now let us insert a known pilot signal $p[n]$ near the source of interest. Denoting the linear convolution sum by $\ast$, we have

$$\mathbf{x}[n] = (\mathbf{a} \ast s)[n] + (\mathbf{a}_p \ast a_t \ast p)[n] + \mathbf{z}[n], \quad (3)$$

where $\mathbf{a}_p \approx \mathbf{a}$ is the AIR vector from the pilot transmitter to the array and $a_t$ is the impulse response associated with the transducer of the transmitter.

For many applications, it is not necessary to measure the transducer response. If the goal of the audio enhancement system is to reduce noise and not to remove reverberation from the source signal, then we need only measure the relative impulse responses (RIR) that relate the microphone signals to each other [16]. Let $\tilde{s}[n] = (a_1 \ast s)[n]$ and $\tilde{p}[n] = (a_{p,1} \ast a_t \ast p)[n]$ be the responses of the first microphone to the source of interest and the pilot signal, respectively. Then, assuming that those impulse responses are invertible, we can write the mixture in terms of the RIRs $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{a}}_p \approx \tilde{\mathbf{a}}$:

$$\mathbf{x}[n] = (\tilde{\mathbf{a}} \ast \tilde{s})[n] + (\tilde{\mathbf{a}}_p \ast \tilde{p})[n] + \mathbf{z}[n], \quad (4)$$

$$\approx (\tilde{\mathbf{a}} \ast (\tilde{s} \ast \tilde{p}))[n] + \mathbf{z}[n]. \quad (5)$$

These relative impulse responses are noncausal in general. In blind audio enhancement algorithms, they can be estimated from the recordings by comparing the frequency-domain magnitudes and phases of the data between channels or by computing a frequency-domain sample covariance matrix [17]. Here, we infer the RIRs from the pilot signal $p[n]$. 

Figure 2. A pilot-assisted audio enhancement system measures the acoustic channel using a known signal emitted near the source of interest.
B. Pilot signals

There are several types of audio signal that are commonly used to measure AIRs. In this proof-of-concept study, we restrict our attention to linear frequency sweeps, which are easy to generate and measure and can be readily distinguished from speech and noise sources in spectrogram representations. Depending on the application, other pilot signals may be more suitable. Logarithmic frequency sweeps are more robust against transducer nonlinearities [18], [19], while pseudorandom noise is less obtrusive to listeners. Certain pseudorandom noise signals, known as maximum length sequences [20], have mathematical properties that make them particularly easy to generate and to invert.

Because linear sweeps have uniform energy at all frequencies of interest, we can measure the AIR up to a scale factor using a cross-correlation:

\[ \hat{a}[n] = \sum_k x[n + k]p[k]. \]  

(6)

The accuracy of the estimate depends on the length of the pilot signal and the noise level in the recording. We will evaluate estimation performance experimentally in Section III-A.

C. Multichannel filter

Once we have estimated the AIRs or RIRs, we can select the filter coefficients \( w[n] \). There are many criteria that can be used to design the filters, depending on our objectives [1]. The filters can be designed to minimize overall mean squared error with respect to the target source, to fully invert the channel and remove reverberation [21], [22], to produce a new mixture of multiple sources [23], [24], or to trade off noise reduction for spectral distortion for each source [25]. Here, we will use the minimum variance distortionless response (MVDR) filter [26], which solves the optimization problem

\[ \min_w \mathbb{E} \left[ \left( w^T z \right)n \right] \]  

s.t. \( w^T \hat{a} \) \( n \) = \( \delta[n] \).  

(7)

That is, it minimizes the output noise power subject to the constraint that signals from the target location must pass through unmodified.

III. EXPERIMENTS

A. Impulse response estimation

First, we consider the accuracy with which we can estimate acoustic impulse responses in noisy and reverberant environments. To quantify the AIR estimation accuracy, we must use the fully synthetic mixture method. We used measured impulse responses from a database of behind-the-ear hearing aid responses in different rooms to simulate mixtures of several speech signals from the TIMIT dataset [27]. The pilot signal was a 100 millisecond linear frequency sweep co-located with one of the speech sources. Figure 3 compares the mean square error between the estimated and ground truth AIRs. At high signal-to-noise ratios (SNR), it is easiest to estimate the anechoic AIRs. The cafeteria, with a reverberation time of \( T_{60} = 1250 \) ms, is more difficult. However, the error is largest for the office. Although the reverberation time is only 300 ms, the source in the office is pointed toward a wall rather than the recording devices, so there is no strong direct path component in the AIR. Sources like this one do not have a well-defined direction of arrival, so it is important to accurately characterize their AIRs or RIRs when designing multichannel filters.

B. Enhancement performance with synthetic mixtures

Next, we evaluate the noise reduction performance of the system. In this experiment, we used the same synthetic hearing aid mixtures from the office environment. The RIRs and the noise statistics were estimated from the noisy mixtures and used to design an MVDR filter. To evaluate the effect of model complexity on performance, the RIRs were truncated using Hamming windows of varying lengths centered at zero lag. Figure 4 shows noise reduction performance as a function of the estimated RIR length. The ground truth filter performed best, followed by the filter designed using a 10 second sweep. The 100 millisecond sweep did not perform as well, but was still better than the anechoic model. There appears to be little benefit to using long RIRs based on a short pilot signal.

C. Enhancement performance with separate recordings

The purpose of the pilot signal method is to enable audio enhancement experiments using real-world recordings. To demonstrate the utility of the method, we recorded speech mixture data using the setup shown in Figure 5. The target speech and pilot signals were generated by two different transducers attached to the neck of a mannequin in a basement storage room with no acoustic treatment. The target speech was a ten second clip of a male TIMIT talker. The background noise signals consisted of different TIMIT talkers played through several arbitrarily placed loudspeakers (not pictured) at around 0 dB SNR. The audio was recorded at a sample rate of 32 kHz by a ReSpeaker Microphone Array, which consists of seven digital MEMS microphones spaced 32 millimeters apart in a hexagonal configuration.

1For the noise statistics, we used a diagonally-loaded sample cross-correlation of the full mixture, so it is technically an MPDR filter [26].
This experiment was conducted using separate source and noise recordings so that the noise reduction performance of the pilot-assisted filter could be measured quantitatively. The noise statistics were estimated using a different segment of speech from the same background talkers. The RIRs were estimated from the pilot, speech, and noise mixture and windowed to 32 ms. Three MVDR filters were designed using:

- RIRs measured in quiet using a ten second linear sweep;
- RIRs estimated during noisy speech using one 100-millisecond pilot every second for ten seconds; and
- an anechoic model based on time differences of arrival.

The gain applied to each signal component is shown in Figure 6. The filter designed using the accurately measured RIRs performed best, applying little distortion to either the pilot or the target speech, and attenuating the background noise by nearly 10 dB. The filter designed using the estimated RIRs did not perfectly match the pilot or target speech signals, and applied slightly less attenuation to the background noise. The similar gains applied to the pilot signal and target speech suggest that the MVDR filter designed for the pilot signal is a close approximation of the MVDR filter for the speech source, even though the two signals were produced by different transducers. The filter designed using the anechoic model significantly attenuated all three signals as it attempted to enhance only the direct path.

D. Enhancement of simultaneous recordings

Finally, we performed an experiment using a simultaneous recording of speech, pilot, and noise signals. The speech data consisted of ten seconds each of a male and female talker from the TIMIT dataset. The pilot signal was a 100 millisecond linear sweep repeated once per second, or twenty times total, with average power comparable to that of the speech signal. The noise signals were pseudorandom “white” noise played through low-quality loudspeakers with a prominent peak near 1.8 kHz. The noise statistics were estimated using the sample cross-correlation of ten seconds of noise data alone. The RIRs were estimated using the twenty linear sweeps, which were recorded simultaneously with the speech and noise signals. Because there is no ground truth data available, it is impossible to quantify the separation performance of this system. However, the spectrograms in Figure 7 illustrate that the fully blind system can achieve substantial noise reduction.

IV. CONCLUSIONS

The experimental results presented here suggest that pilot signals placed near a speech source can be used to measure the acoustic channel between the source and a microphone array. Filters designed using these reverberant channel measurements can outperform conventional anechoic beamformers, especially when the signal of interest does not have a strong direct path to the array. Our experiments suggest that useful relative impulse response measurements can be obtained even with short pilot signals in strong background noise. Furthermore, we were able to achieve significant noise reduction in a real recording using low-quality, uncalibrated transducers and microphones in an untreated room. The pilot signals let us design useful filters with no prior information on the room geometry, array configuration, or signal characteristics. Thus, the pilot method allows us to evaluate the achievable performance of multichannel audio enhancement systems in uncontrolled real-world environments.
Figure 7. Spectrograms of part of a real-world speech-and-noise mixture with 100 ms pilots.

REFERENCES


