Audiovisual Zooming: What You See Is What You Hear

Arun Asokan Nair Austin Reiter Changxi Zheng Shree Nayar
Snap Research

ABSTRACT

When capturing videos on a mobile platform, often the target of interest is contaminated by the surrounding environment. To alleviate the visual irrelevance, camera panning and zooming provide the means to isolate a desired field of view (FOV). However, the captured audio is still contaminated by signals outside the FOV. This effect is unnatural—for human perception, visual and auditory cues must go hand-in-hand.

We present the concept of *Audiovisual Zooming*, whereby an auditory FOV is formed to match the visual. Our framework is built around the classic idea of beamforming, a computational approach to enhancing sound from a single direction using a microphone array. Yet, beamforming on its own can not incorporate the auditory FOV, as the FOV may include an arbitrary number of directional sources. We formulate our audiovisual zooming as a generalized eigenvalue problem and propose an algorithm for efficient computation on mobile platforms. To inform the algorithmic and physical implementation, we offer a theoretical analysis of our algorithmic components as well as numerical studies for understanding various design choices of microphone arrays. Finally, we demonstrate audiovisual zooming on two different mobile platforms: a mobile smartphone and a 360° spherical imaging system for video conference settings.

CCS CONCEPTS

• **Information systems** → *Multimedia content creation.*

KEYWORDS

audiovisual zooming, beamforming, audio enhancement

ACM Reference Format:
Arun Asokan Nair Austin Reiter Changxi Zheng Shree Nayar. 2019. Audiovisual Zooming: What You See Is What You Hear. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3343031.3351010

This work was done at Snap's NYC Research Lab and supported by Snap Inc. Arun Asokan Nair is with the Department of Electrical and Computer Engineering at Johns Hopkins University e-mail:anair8@jhu.edu. Austin Reiter is with the Applied Machine Learning group at Facebook. Changxi Zheng is with the Department of Computer Science at Columbia University.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351010

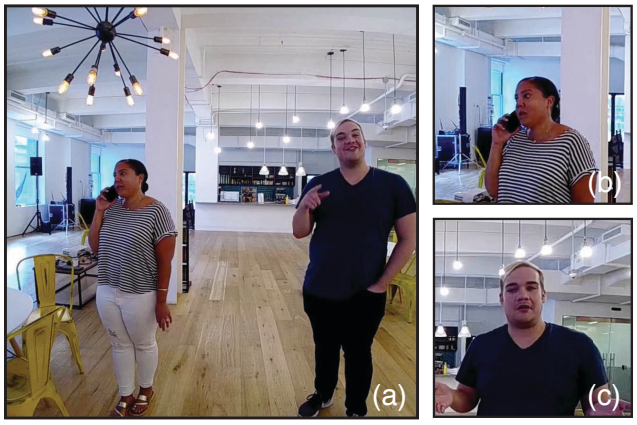

Figure 1: Audiovisual zooming. When the camera captures both people (a), we hear them both talk. (b) As the camera zooms in and focuses on the woman, her speech in the captured video is enhanced while the man's speech is suppressed. (c) Then, the camera pans and focuses on the man, in this process his speech becomes more pronounced while the woman's speech fades out. In our system, the camera's FOV synchronizes with its auditory focus—what you see is what you hear (see supplementary video).

1 INTRODUCTION

The camera can tilt, pan, pedestal, dolly, truck, and zoom—to control what the viewer sees. Historically, this rich vocabulary of camera control is only at the professional's disposal. Today, every mobile device is equipped with a compact and light camera, allowing anyone to decide what imagery in what way is to be captured. Whenever one captures a video, audio is also captured, but the vocabulary with which a user can exert control over the audio pales in comparison to user control over the video. No matter where the camera is pointed or how zoomed it is, the sound is always recorded regardless of its incoming direction, be it from behind the camera or somewhere in the view. As a result, the captured video might not match the audio, leading to an unnatural experience.

The problem is that the camera lacks an *auditory field of view*, one that is synchronized with and driven by the camera's optical field of view (FOV). In this work, we introduce the concept of focusing an auditory FOV (see Figure 1) to address the problem. We call our concept **Audiovisual Zooming**.

The closest field-of-study to this concept is *Beamforming* [9], a computational technique that constructs a directional microphone by using an array of omnidirectional microphones. Leveraging the different time delays of signals that arrive from different directions, the idea is to linearly combine microphone signals into an output signal boosting the sound coming from a *target direction*, while suppressing everything else for monoaural directional sound filtering. In almost all beamforming techniques, the single target direction

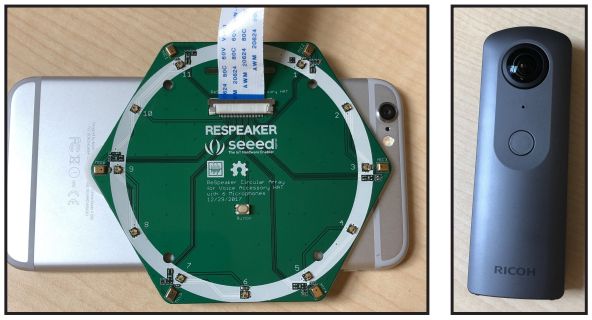

Figure 2: Audiovisual zooming is implemented on two mobile platforms: an off-the-shelf planar microphone array (with 6 microphones) is attached to a smartphone [left] and a 360° Ricoh camera [right] to show smartphone and teleconferencing utility.

needs to be specified or estimated, and plays an important role in the mathematical formulation of beamforming. It is this essential notion of target direction that sets apart our method from the traditional beamforming.

Our audiovisual zooming requires no target direction. In contrast, we introduce auditory FOV, which defines a directional region (i.e., a solid angle area) consistent with the camera's optical FOV. All sounds, no matter how many, coming from within this region are enhanced, while those outside of the region are suppressed. In this way, the captured audio is in synchronization with the captured imagery. In other words, *what you see is what you hear*.

One approach toward this goal is jointly analyzing the captured audio and visual content through deep learning [7, 23, 37]. The success of this approach lies in the strong correlation between the motion in captured imagery and the resulting audio, as well as the feasibility of constructing a large training dataset. But often the motion-audio correlation is weak or even undetectable—for example, when the sound source is far from the camera, or occluded by other objects (but still in the FOV). In addition, there may be arbitrary numbers/types of sound sources in the FOV. Constructing a training dataset that covers all these cases quickly becomes intractable, and the resulting deep neural networks are unlikely to run on a low-budget mobile device where the camera often resides.

Technical contributions. In this work, we augment the microphone array and beamforming approaches to enable audiovisual zooming, without learning from training data. Motivated by microphone array beamforming, we view the signals sampled by individual microphones as random variables of some underlying stochastic process. From this perspective, we estimate two complex-valued matrices, called *spectral matrices*, in frequency domain: one describes the autocorrelation and cross-correlation of microphone signals that come from within the FOV, and the other describes signals coming from outside of the FOV. We show that with these two matrices, the problem of enhancing towards an FOV can be formulated as a generalized eigenvalue problem that can be easily solved on a mobile device. Our approach is not meant to improve beamforming, but rather to enable audiovisual zooming.

To analyze our approach, we derive a theoretical error bound for our spectral matrix estimation, and reveal a connection of the error residual to the performance of the classic minimum variance distortionless response (MVDR) beamformer. Empirically, we conduct simulations to understand how various design parameters affect a microphone array.

These inferences inform our implementation. Our final algorithm is simple and can be easily deployed on mobile devices. We realize the audiovisual zooming system by attaching a planar microphone array to two different mobile imaging platforms: a mobile smartphone and a 360° spherical imaging system for teleconference settings (see Figure 2). Finally, we demonstrate our system in a number of use cases.

2 RELATED WORK

Our audiovisual zooming is built on classic beamforming. We therefore briefly review related work in this area. We also discuss the difference of our approach from the general idea of audiovisual machine learning approaches.

Beamforming. A rich and mature research field, acoustic beamforming has a long history, dating back to 1970s when Billingsley [3] invented the microphone antenna called the acoustic telescope. We refer the reader to [20] for a review of the development of acoustic beamforming techniques and to [9] for an exhaustive survey of the state of the field.

In general, the various beamforming methods falls into one of two categories: fixed and adaptive. Fixed beamformers are best summarized by the well-known *Delay-and-Sum* method [30, 33], which delays the signal received by each microphone according to the relative propagation delays from a target direction, and then sums the signals together across the microphones. This serves to enhance the gain of the target direction, but often does little to suppress anything else.

The seminal work of Capon [5] introduces an adaptive, or *data-dependent*, beamforming technique, later known as the Minimum Variance Distortionless Response (MVDR) beamformer [28, 32]. This approach optimizes a set of weights to linearly combine the signals in time-frequency space so as to minimize residual noise and constrain the sound from the desired direction to be undistorted. The robustness of MVDR beamformer is later improved by various extensions such as dynamic loading [18]. Since our method is built on the MVDR beamformer, we will briefly review its formulation in §3.1. There are also other variants, such as the Linearly Constrained Minimum Variance (LCMV) [10], Principal component [14, 36], and Generalized Eigenvalue [35] beamformers, all of which are special cases of a shared underlying optimization framework [32].

All these beamforming techniques have the same goal: enhancing the sound from *a single direction*, and they have no notion of *field of view* (FOV). In contrast, our goal is to enhance all sounds from within an *arbitrary FOV* and suppress everything outside. It is this very difference that requires a different beamforming formulation and thus necessitates the development of a new algorithm.

Recently, a few methods have been proposed to enhance sounds from multiple sources. Thiergart et al. [31] introduced acoustic zoom, wherein all detected sound sources are individually isolated via direction-of-arrival (DOA) estimation and beamforming, and

then combined through a weighting scheme defined by their zooming parameters. Ruochen et al. [27] used a spherical microphone array and the psychoacoustic theory to model sound perceptions and control audio boosting using camera metadata in so-called B-Format Encoding. A more recent method [6] uses MVDR beamforming in three orthogonal directions and chooses the sound from the microphone closest to the target region. Just by choosing a microphone signal, this method does not enhance received sound, and is inherently limited for small form-factor arrays. Our method, in contrast, requires no estimation of DOAs and can be implemented using compact microphone arrays.

Audiovisual learning. Recently, a line of work has emerged that combines computer vision and audio via deep learning for speech recognition, separation, and enhancement [8, 12, 21, 25]. Particularly related to our work, Ephrat et al. [7] recently introduced a deep learning model that detects and analyzes facial movements along with learning a mask on Fourier coefficients to mask out desired speech associated with particular facial motion. Zhao et al. [37] addressed a similar problem of separating the sound of multiple on-screen objects by training a self-supervised model. Owens and Efros [23] used a deep neural network to predict whether audio and visual tracks are temporally aligned. Features learned through training are then used to perform an on/off screen speaker separation. Afouras et al. [2] trained a deep neural network that takes audio and visual cues to denoise speech spectrograms. While impressive, these work require that the visual component of the sound is both visible and has sufficient pixel resolution to capture the appearance and motion. Our work does not rely on any analysis of visual cues, and as such, can enhance sound coming from any FOV even when the motion that produces this sound is occluded or far away from the camera.

Summary. Our method differs from previous works in that 1) no knowledge of DOAs is required, 2) the user may specify any arbitrary FOV to match that of a camera's, and 3) our approach will enhance only the sound from within that FOV and attenuate everything else. *In this way, the camera drives the experience entirely, forcing the focused audio content to match what is being viewed.*

3 THEORY OF AUDIOVISUAL ZOOMING

A cornerstone of our audiovisual zooming system is microphone array beamforming. To understand our algorithm, we start with a brief review of this classic technique.

Microphone array model. We consider a microphone array that consists of M sensors receiving sound from all directions. The time-domain signals captured by microphone i ($i = 1 \ldots M$) is

$$y_i(t) = \sum_{s=1}^{S} h_{s \to i}(t) * x_s(t) + n_i(t), \tag{1}$$

where $*$ denotes the convolution operator, s indices individual sound sources, $x_s(t)$ is the signals emitted at sound source s, $n_i(t)$ is the noise at microphone i, and $h_{s \to i}(t)$ is the *Acoustic Transfer Function* for source s impinging on microphone i. This transfer function accounts for how the sound propagates from s to i, including both direct and indirect propagation (e.g., reflection and diffraction by the environment).

Because the sound propagation largely depends on its frequency components, the microphone array model is often expressed in time-frequency (T-F) domain [9] through the Short-Time Fourier Transform (STFT). In T-F domain, the convolution operator becomes into a multiplication, and Eq. (1) is written as

$$Y_i(n, \omega) = \sum_{s=1}^{S} H_{s \to i}(n, \omega) X_s(n, \omega) + N_i(n, \omega), \tag{2}$$

where n and ω index the time frame and the discrete frequency bin, respectively. We then stack the STFT coefficients for all sensors in a vector,

$$\boldsymbol{Y}(n, \omega) = [Y_1(n, \omega), \ldots, Y_M(n, \omega)]^T. \tag{3}$$

With these notations, we now briefly review the classic beamforming algorithms, as follows.

3.1 Beamforming Briefing

The general idea of beamforming is simple. It linearly combines the input multi-channel signals into a mono-channel signal in T-F domain. Provided a set of frequency-dependent weights $\boldsymbol{w}(\omega) = [w_1(\omega), \ldots, w_M(\omega)]^T$, the linear combination outputs a signal as $\boldsymbol{w}^H(\omega)\boldsymbol{Y}(n, \omega)$, where the superscript H denotes conjugate transpose. By carefully choosing the weights $\boldsymbol{w}$, the resulting signal enhances the sound received from a given *single* direction $\boldsymbol{d}$.

Intuitively, this is possible because the sound signals recorded at different microphones differ in both amplitude and phase. One can choose the weights $\boldsymbol{w}$ to "adjust" the differences such that when the signals are superimposed, they interfere constructively for sound coming from the direction $\boldsymbol{d}$ but destructively for sound from other directions. Numerous algorithms have been devised to estimate the weights $\boldsymbol{w}$. Here we only review the ones that are most relevant to our method, while referring the reader to the textbooks [4, 32] for a comprehensive introduction.

3.1.1 Spectral matrix. A fundamental philosophy in microphone array processing is to model the received signal as a *stochastic process*. Each individual sample $y_i[t]$ of microphone i is assumed to be an outcome of some underlying random process.

An important notion from this vantage point is the *spectral matrix*, an $M \times M$ complex-valued Hermitian matrix, denoted as $\mathrm{R}(\omega)$, describing the frequency-domain signal statistics received by the microphone array. Its diagonal element $\mathrm{R}_{ii}(\omega)$ indicates the *auto-correlation* (in frequency domain) of the impinging signal received by microphone i, that is, the power spectrum of the signal at i. Its off-diagonal element $\mathrm{R}_{ij}(\omega)$ describes the *cross-correlation* of signals received by microphone i and j, reflecting the phase differences between the two microphone signals. In short, the spectral matrix encapsulates information needed for the estimation of $\boldsymbol{w}$—toward constructively enhancing the signal of a given direction.

In practice, $\mathrm{R}(\omega)$ is estimated using the frequency-domain snapshots $\boldsymbol{Y}(n, \omega)$ in (3). A simple yet common estimator is

$$\mathrm{R}(\omega) \approx \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{Y}(n, \omega) \boldsymbol{Y}^H(n, \omega), \tag{4}$$

from which many improvements have been developed (such as the Forward-Backward averaging [28]).

If a set of weights $\boldsymbol{w}(\omega)$ is used to combine the microphone signals in the frequency band ω, it can be shown that the output signal has the power spectrum expressed as $\boldsymbol{w}^H(\omega)\mathrm{R}(\omega)\boldsymbol{w}(\omega)$ [4]). From now on, when there is no confusion, we will ignore the frequency parameter ω and simply write R and $\boldsymbol{w}$.

3.1.2 Minimum Variance Distortionless Response (MVDR) beamformer. In beamforming theory, the spectral matrix R is viewed as a composition of two parts, signal spectral matrix $\mathrm{R_s}$ and noise spectral matrix $\mathrm{R_n}$. $\mathrm{R_s}$ accounts for the signal solely from the desired direction $\boldsymbol{d}$ (sometimes also called the direction of arrival), while $\mathrm{R_n}$ accounts for the *unwanted signals* including both the ambient noise (i.e., N_i in (2)) and those from the undesired directions. Note that R might not be a simple summation of $\mathrm{R_s}$ and $\mathrm{R_n}$ if the unwanted signals and the desired signal are (at least partially) correlated.

The classic MVDR beamformer finds the optimal $\boldsymbol{w}$ in the following sense: it minimizes the power of unwanted signals, while keeping the signal from the desired direction undistorted. This is formally expressed as a constrained optimization problem,

$$\boldsymbol{w}_{\mathrm{BF}} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^H \mathrm{R_n} \mathbf{w}, \quad \text{s.t. } \boldsymbol{w}^H \boldsymbol{v}_d = 1. \tag{5}$$

Here the objective function measures the power of unwanted signals in the output. The constraint requires the incoming signal from the direction $\boldsymbol{d}$ to remain undistorted in the output signal. The vector $\boldsymbol{v}_d$, called the *steering vector*, indicates the relative phases of the signal impinging on all M microphones from the desired direction $\boldsymbol{d}$, defined as

$$\boldsymbol{v}_d = \left[e^{-j\frac{\omega}{c}\boldsymbol{d}^T\boldsymbol{p}_1} \quad \dots \quad e^{-j\frac{\omega}{c}\boldsymbol{d}^T\boldsymbol{p}_M} \right]^T, \tag{6}$$

where c is the speed of sound, and $\boldsymbol{p}_i$ ($i = 1 \dots M$) is the spatial position of each microphone in the array. The steering vector describes the relative phase difference of a plane wave sound impinging on the microphones from direction $\boldsymbol{d}$. The intuition behind the constraint in (5) is that the weights $\boldsymbol{w}$ need to compensate the received phase differences at the microphones from direction $\boldsymbol{d}$ and thereby constructively combine the signals to boost the signal from $\boldsymbol{d}$.

MVDR beamformer is one of the most widely used beamforming techniques. Provided a single steering direction $\boldsymbol{d}$ and an estimation of $\mathrm{R_n}$, it has realtime performance even on a low-budget mobile device, since the weights can be analytically written as

$$\boldsymbol{w}_{\mathrm{BF}} = \frac{\boldsymbol{v}_d^H \mathrm{R_n^{-1}}}{\boldsymbol{v}_d^H \mathrm{R_n^{-1}} \boldsymbol{v}_d}. \tag{7}$$

Oftentimes, however, estimation of $\mathrm{R_n}$ from microphone recordings is challenging. An approximation is by replacing $\mathrm{R_n}$ in (5) with the spectral matrix R, as R can be directly estimated using the recorded signals in (4). Then, the optimization objective is to minimize the total output power subject to the constraint in (5). This is the so-called *Minimum Power Distortionless Response* (MPDR) beamformer, one that lays the foundation of our audiovisual zooming method.

3.2 Beamforming Toward a Field of View

Almost all beamforming techniques require to know a steering direction $\boldsymbol{d}$. Indeed, this single steering direction is pivotal for establishing the constraint in MVDR/MPDR formulation (5). However, in our work, we wish to enhance signals toward a field of view (FOV), that is, a *continuous set* of steering directions (Figure 3).

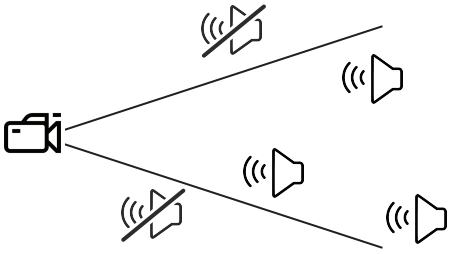

Figure 3: Unlike traditional beamforming, our audiovisual zooming system does not rely on a specific target direction. Any sound sources captured by the camera's FOV will be enhanced, while those outside of the FOV are suppressed.

How to incorporate the FOV in microphone array beamforming is the challenge that we need to overcome.

We determine the beamforming FOV based on the camera's FOV (elaborated in §5.3). We also note that the beamforming FOV may vary as the user zooms in/out or pans the camera.

3.2.1 Generalized eigenvalue formulation. The starting point of our method is also the spectral matrix (recall §3.1.1). Yet, for our purpose of beamforing toward an FOV, the signal and noise spectral matrices, $\mathrm{R_s}$ and $\mathrm{R_n}$, must be interpreted in a different way. Now, $\mathrm{R_s}$ accounts for all signals coming from directions inside the FOV, while $\mathrm{R_n}$ includes signals outside of the FOV. Suppose for now we know both $\mathrm{R_s}$ and $\mathrm{R_n}$. We can formulate a beamforming optimization problem by maximizing the output signal-to-noise (SNR) ratio, namely

$$\boldsymbol{w}_{\mathrm{FOV}} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^H \mathrm{R_s} \boldsymbol{w}}{\boldsymbol{w}^H \mathrm{R_n} \boldsymbol{w}}. \tag{8}$$

Here the numerator and denominator measure the powers of desired signals and unwanted signals, respectively. This formulation is known in traditional beamforming, although not widely used. This is because it needs the estimation of both $\mathrm{R_s}$ and $\mathrm{R_n}$, and when a single steering direction is considered (e.g., when the desired signal is a plane wave along a direction), this formulation is identical to MVDR beamformer [32]—no need to solve (8) directly.

But this formulation is significant for our problem, since it requires no steering direction. Indeed, the desire of steering toward an FOV can be expressed by $\mathrm{R_s}$, which can include signals from an arbitrary set of directions. If $\mathrm{R_s}$ and $\mathrm{R_n}$ can be robustly estimated, then solving for $\boldsymbol{w}_{\mathrm{FOV}}$ amounts to a simple generalized eigenvalue problem (by noticing that the objective in (8) is a generalized Rayleigh quotient):

$$\mathrm{R_s} \boldsymbol{w} = \lambda \mathrm{R_n} \boldsymbol{w}. \tag{9}$$

The solution $\boldsymbol{w}_{\mathrm{FOV}}$ in (8) is the eigenvector of the maximal eigenvalue.

3.2.2 Estimation of signal and noise spectral matrices. The remaining question is how to estimate $\mathrm{R_s}$ and $\mathrm{R_n}$ that respect the FOV. Some traditional beamforming methods (such as MVDR) also need to estimate $\mathrm{R_n}$, for which a popular approach is by estimating a noise mask in T-F domain [13]. There, a common assumption is that the desired signal is the speech of a single voice. In other words, it assumes that the desired signal has a T-F structure, which can

be inferred and used to estimate the mask by a machine learning model trained over a large speech dataset.

In our problem, the desired signals are those received in the FOV. In stark contrast to the single speech assumption, their structure is unclear, as they might include an arbitrary number of speakers, other types of sound, and even ambient noise coming from the FOV. It is too expensive to construct a sufficient training dataset for a machine learning model producing reasonable masks.

We resort to the MPDR beamformer to estimate R_s and R_n. First, consider a direction $\boldsymbol{\theta}$. The MPDR weights $\boldsymbol{w}_\theta$ for enhancing signals from $\boldsymbol{\theta}$ is expressed in (7), where the steering vector $\boldsymbol{v}_d = \boldsymbol{v}_\theta$ is defined in (6) and the matrix R_n is replaced by the total spectral matrix R estimated using (4). Substituting this expression in $\boldsymbol{w}_\theta^H \mathsf{R} \boldsymbol{w}_\theta$ yields the power spectrum of the MPDR output signal,

$$P(\boldsymbol{\theta}) = \left[\boldsymbol{v}_\theta^H \mathsf{R}^{-1} \boldsymbol{v}_\theta\right]^{-1}. \tag{10}$$

Recall that the effect of MPDR beamformer is to boost the signal from direction $\boldsymbol{\theta}$ while suppressing signals from other directions. Thus, $P(\boldsymbol{\theta})$ can be viewed as an estimation of the power of a plane wave coming from the direction $\boldsymbol{\theta}$.

Using this estimation, the microphone array spectral matrix resulted from the sound wave *only* from $\boldsymbol{\theta}$ direction is written as $P(\boldsymbol{\theta})\boldsymbol{v}_\theta \boldsymbol{v}_\theta^H$ (see Appendix A for more details). If we assume that signals from different directions are uncorrelated, then the signal spectral matrix for sound coming from an FOV is an integral of the single-direction estimation over the entire FOV:

$$\mathsf{R}_\mathrm{s} \approx \int_{\boldsymbol{\Theta}} P(\boldsymbol{\theta}) \boldsymbol{v}_\theta \boldsymbol{v}_\theta^H \mathrm{d}\boldsymbol{\theta}, \tag{11}$$

where $\boldsymbol{\Theta}$ is the solid angle area spanned by the camera's FOV, set by the current camera direction and zoom settings. Similarly, the noise spectral matrix R_n can be estimated using the same integral but over the solid angle area $\mathcal{S}_3 \setminus \boldsymbol{\Theta}$, where $\mathcal{S}_3$ denotes the solid angle of an entire 3D sphere. Note that both R_s and R_n are frequency dependent, and thus they are estimated for each individual frequency band. The estimation (11) can be applied to an arbitrary FOV, and is agnostic to the sound source distribution in the FOV.

We note that a similar integral has been used for standard MVDR beamformer [11] to estimate the spectral matrix excluding a single sound direction. However, the accuracy of the power spectrum estimation (10) and the matrix estimation (11) are unclear. In the rest of this section, we theoretically analyze and justify this estimation.

3.2.3 Analysis. Our estimation of the single-direction power $P(\boldsymbol{\theta})$ is built on the MPDR beamformer. Traditionally, a beamformer is used to enhance sound from a direction $\boldsymbol{s}$. It has been shown that the MPDR beamformer is identical to the MVDR beamformer when the steering direction $\boldsymbol{d}$ (in MPDR) is chosen to be the true sound source direction $\boldsymbol{s}$, but MPDR beamformer is much less reliable [4]: a small mismatch between $\boldsymbol{d}$ and $\boldsymbol{s}$ can degrade significantly the MPDR performance. Fortunately, this is not an issue in our case, since we have no explicit notion of sound sources. When evaluating the integral (11), we treat each direction $\boldsymbol{d}$ in the FOV as a true sound source direction $\boldsymbol{s}$.

Next, we present an analytical understanding of (10) and (11) for spectral matrix estimation. First, if the recording environment has only (uncorrelated) ambient noise, then the acoustic power is uniform in space, and the spectral matrix R has the form, $\mathsf{R} = \sigma \mathsf{I}_M$, where σ is the ambient noise power, and I_M is an $M \times M$ identity matrix, where M is the number of microphones. We therefore expect the estimated R_s to have power proportional to the FOV area . In this case, $P(\boldsymbol{\theta})$ is a constant σ/M, and R_s in (11) indeed has diagonal elements proportional to the FOV area. Now, consider the general case of estimating the signal power $P(\boldsymbol{\theta})$ for the direction $\boldsymbol{\theta}$. Assuming signals from different directions are uncorrelated, then the (true) spectral matrix can be decomposed into two,

$$\mathsf{R} = \mathsf{R}_\mathrm{c} + m_\theta \boldsymbol{v}_\theta \boldsymbol{v}_\theta^H, \tag{12}$$

where R_c accounts for the sound signals from all directions but $\boldsymbol{\theta}$, and the second term is the contribution of a plane wave coming from $\boldsymbol{\theta}$: m_θ is its power, and $\boldsymbol{v}_\theta$ is a vector defined in (6) (see Appendix A for more explanation of the plane wave contribution). To see how well the estimation (10) approximates the true power m_θ, we express $P(\boldsymbol{\theta})$ analytically by applying the matrix inversion lemma [19] on R and obtain

$$\begin{aligned} P(\boldsymbol{\theta}) &= \left[\boldsymbol{v}_\theta^H \mathsf{R}_\mathrm{c}^{-1} \boldsymbol{v}_\theta - \frac{m_\theta}{1 + m_\theta \boldsymbol{v}_\theta^H \mathsf{R}_\mathrm{c}^{-1} \boldsymbol{v}_\theta} (\boldsymbol{v}_\theta^H \mathsf{R}_\mathrm{c}^{-1} \boldsymbol{v}_\theta)^2\right]^{-1} \\ &= \left(a - \frac{m_\theta a^2}{1 + m_\theta a}\right)^{-1} \\ &= m_\theta + \frac{1}{a}, \end{aligned} \tag{13}$$

where a denotes $\boldsymbol{v}_\theta^H \mathsf{R}_\mathrm{c}^{-1} \boldsymbol{v}_\theta$ for short. It is evident the estimated power $P(\boldsymbol{\theta})$ differs from the true power m_θ by a constant $1/a$. In fact, $1/a$ is the noise power in the output signal from the MVDR beamformer (i.e., using R_c in (5) and computing $\boldsymbol{w}_\mathrm{FB}^H \mathsf{R}_\mathrm{c} \boldsymbol{w}_\mathrm{FB}$), and the MVDR beamformer is designed to minimize exactly this noise power ($1/a$). Here noise is all the signals *not* from direction $\boldsymbol{\theta}$. Thus, from (13), we conclude that the estimation accuracy of $P(\boldsymbol{\theta})$ depends on $m_\theta a$, the output SNR ratio of the MVDR beamformer.

Furthermore, we show that the estimation of R_s in (11) has a bounded error. Formally, we rewrite (11) as

$$\mathsf{R}_\mathrm{s} = \int_{\boldsymbol{\Theta}} m_\theta \boldsymbol{v}_\theta \boldsymbol{v}_\theta^H \mathrm{d}\boldsymbol{\theta} + \Delta. \tag{14}$$

The first term here is the true signal spectral matrix, and Δ is the error residual introduced by the estimator (11). As derived in Appendix B, Δ is bounded from above and below:

$$\frac{\lambda_\mathrm{min}}{M} \int_{\boldsymbol{\Theta}} \boldsymbol{v}_\theta \boldsymbol{v}_\theta^H \mathrm{d}\boldsymbol{\theta} \le \Delta \le \frac{\lambda_\mathrm{max}}{M} \int_{\boldsymbol{\Theta}} \boldsymbol{v}_\theta \boldsymbol{v}_\theta^H \mathrm{d}\boldsymbol{\theta}, \tag{15}$$

where λ_min and λ_max are the minimal and maximal eigenvalues of R_c, respectively. This derivation indicates that the residual is bounded from above, proportional to the power of the strongest signal direction other than $\boldsymbol{\theta}$ and inversely proportional to the microphone array size.

In light of this, a simple strategy for improving estimation accuracy of R_s is improving the MVDR's output SNR ratio or increasing the number of microphones (i.e., M). In the next section, we provide more guidance on microphone array design for audiovisual zooming through numerical simulations.

4 EMPIRICAL STUDIES OF ARRAY DESIGNS

We implement our audiovisual zooming system on a microphone array, which has many design parameters. Yet, there is no golden rule to set those parameters; they depend on specific applications [16]. We conduct a series of simulation experiments to understand the design parameters tailored for our applications, wherein the mobile device such as a smartphone is the form factor that we will restrict the microphone array to fit in. Concretely, we explore the following questions:

- How does beamforming change with frequency?
- How big should the array be?
- What number of microphones should we use?
- How should we sample the spatial directions in (11)?

The first three questions are to understand microphone array configuration, while the last is for efficient implementation of our audiovisual zooming algorithm. Because the audiovisual zooming is based on MVDR beamforming, we must understand how the beamforming performance changes with respect to the array's design parameters. Therefore, the studies here are not meant to evaluate our audiovisual zooming method. Rather, we examine MVDR beamforming under different setups to understand design parameter choices. While there have been plenty of empirical studies of microphone array parameters (e.g., see [1, 9, 24]), our primary goal of conducting these studies is to inform our specific algorithm.

In this section, we present the conclusions we learned from the empirical studies. The details of our simulations and their results are in Appendix C of the supplementary.

Setup. We consider a circular array consisting of a number of omnidirectional microphones evenly placed on a circle in the X-Y plane (see Figure 7 in Appendix C) centered at the origin. We choose this configuration because it matches the off-the-shelf physical array that we will use. We also place six sound sources throughout the space: four orthogonal sources in the X-Y plane at 0°, 90°, 180°, and 270°. The other two are placed along the positive and negative Z-axes, respectively. The environment is filled with ambient Gaussian noise. The MVDR beamformer aims to enhance the sound coming from the positive Z-direction, while suppressing everything else.

Frequency dependence. MVDR performance is frequency dependent. Our experiments focus on the frequency range of typical human speech (i.e., 300-3420Hz). The results are visualized as MVDR beam patterns in Figure 7 of Appendix C at 300Hz, 1860Hz, and 3420Hz. Towards the steering direction, the beamformer always has unit gain, thanks to the distortionless constraint in (5), but its shape varies across frequencies. In general, *beamforming performance increases as frequency increases.* The beam pattern at 3420Hz (Figure 7-d) also shows some side lobs near the X-Y plane—a phenomenon known as *spatial aliasing* occurring at high frequencies.

Array size. Next, we study the effect of the overall size of the array: the number of microphones is fixed and the inter-microphone spacing is changed. Figure 8 in Appendix C shows the details of our studies. In general, the simulations show that *better directionality requires a larger array size*—but not too large. For example, once the array size reaches 50cm, we get *non-trivial spatial aliasing effects.* Though there is a strong gain toward the target direction, there are also many unwanted secondary gains in other directions. One way to avoid spatial aliasing is to increase the spatial sampling rate. This brings up the next natural question: how many microphones should we use?

Number of microphones. We simulate the beamformer response as we change the number of microphones while fixing the overall size (i.e., 5cm in radius). The results are shown in Figure 10 in Appendix C. When we increase the number of microphones, we obtain better suppression of the interference relative to the target. However (and perhaps somewhat surprisingly), the performance plateaus once we have sufficient microphones. Figure 10 shows that 16 microphones become superfluous, yielding no improvement over 8. In other words, *more microphones improve performance, but have diminishing returns.*

For the 8- and 16-microphone cases, there are "indentations" in the directions of the X-Y plane interferers (bottom row of Figure 10), indicating so-called *null* responses toward those directions—this is the desired effect. Although there are reasonably larger gains near the areas of those indentations, no sounds comes from those directions in our setup, and so no suppression is needed. This is the advantage of adaptive beamformers (like MVDR): they work to rearrange the gain distribution to best nullify interferers while distributing energy in places where no sound is thought to be.

Sampling density. Our audiovisual zooming algorithm estimates spectral matrices R_n and R_s in (11) through Monte Carlo integration. In practice, we need to sample directions within a desired FOV $\boldsymbol{\Theta}$ and at its outside $\mathcal{S}_3 \setminus \boldsymbol{\Theta}$. The sampling density of the directions should not be arbitrary because for each target we beamform towards, there is an effective main lobe with a non-trivial width, meaning that although the gain in the direction of the target is maximal, there is also non-zero gain from directions nearby the target direction. As shown in Figure 9 in Appendix C, the gain falls off as the sound incoming direction deviates from the target direction. We determine an acceptable reduction in dB (i.e., 1.8dB) for nearby sounds and Figure 9 suggests sampling directions with an angular separation of 20°.

Discussions: extending to 3D arrays. Thus far, most microphone arrays have a 2D planar configuration. We note that there is inherently a symmetry. For the sound wave coming from an elevation angle θ in spherical coordinates, no 2D planar array can disambiguate it from the wave coming from the same but negated elevation angle $-\theta$ (and the same azimuthal angle). For our applications, this is not a significant problem, as the sound waves from behind the array are often blocked by the user who is holding the camera to capture or the table on which the array is placed (see Figure 4). Nevertheless, here we study the performance of a 3D array for future extension. We add an additional microphone at the center of the array and gradually move it along the negative Z-axis to break the planar symmetry. We then examine how this affects the beam pattern. As shown in Figure 11, this additional microphone indeed helps to break the symmetry. As it moves further away from the microphone plane, the interferer behind the array attenuates more. However, such a 3D array is much more bulky than the 2D array. Today, the form factor of a mobile device is one of the most decisive factors for its use on a daily basis. It is unclear if a 3D array is worth equipping on mobile devices.

	MVDR	Our Method
SDR [dB]	-2.96793	**-0.0095**
SNR [dB]	-0.86467	**1.80908**
WADA-SNR [dB]	5.1604	**7.55923**
STOI	0.66414	**0.71992**
PESQ	1.75125	**2.04402**

Table 1: Comparison of our method against MVDR. Our method consistently outperforms MVDR.

5 EXPERIMENTS AND RESULTS

We demonstrate our results via experiments on both synthetic and real data: first, we use synthetic mixtures of clean speech sources in various configurations to evaluate audiovisual zooming enhancement (§5.1) using the following quantitative metrics:

(1) Signal-to-Distortion-Ratio (SDR) [34]. SDR evaluation takes as input the enhanced signal and the reference signal it should ideally match. It first decomposes the enhanced signal into four components: a target component coming from the reference signal, an interferer component containing other unwanted sources' contributions, a noise component encapsulating sensor noises and an artifact component capturing distortions from other sources (like forbidden distortions of the sources and/or "burbling" artifacts). SDR is then calculated as the logarithmic ratio (in dB) of the energy in the target component to the energy in the unwanted components.
(2) Signal-to-Noise-Ratio (SNR). SNR is defined here to be the logarithmic ratio (in dB) of the energy of the enhanced signal to that of the noise signal, the latter of which is defined at each time point as the difference between the enhanced signal and the reference signal.
(3) Waveform Amplitude Distribution Analysis SNR (WADA-SNR) [15]. This metric evaluation assumes that clean speech has an amplitude distribution well approximated by the Gamma distribution with a shaping parameter of 0.4, and that the additive noise signal is Gaussian. It is calculated by studying the amplitude distribution of the enhanced signal. As the reference signal we are attempting to recover in our experiments is oftentimes speech, we use this metric to measure enhancement quality that better correlates with our task.
(4) Short-Time Objective Intelligibility (STOI) [29]. Popular objective measures such as SDR and SNR above often do not reflect well the speech intelligibility–how easily the resulting signals can be understood by humans. The STOI score is designed to bridge that divide.
(5) Perceptual Evaluation of Speech Quality (PESQ) [26]. Similar to STOI, common measures like SDR and SNR do not correlate well with voice quality evaluation results from humans. PESQ was developed to model these subjective tests better, and is a widely used industry standard for objective voice quality testing.

Next, we perform real experiments using audio loud speakers in various configurations to compute SDR, SNR, WADA-SNR, STOI and PESQ enhancements. Finally, we show qualitative performance using two different hardware platforms to demonstrate feasibility in mobile settings. For all experiments, we tested our method against the MVDR beamforming approach as a basis-of-comparison.

Metric	Method	90°	45°	30°	15°
SDR [dB]	MVDR	-4.08	-1.10	-3.54	-1.98
	Ours	-2.11	0.08	-2.56	-1.29
SNR [dB]	MVDR	-0.55	0.78	-0.52	-0.19
	Ours	0.75	1.86	0.48	0.79
WADA-SNR [dB]	MVDR	-0.24	1.73	1.13	1.60
	Ours	1.30	3.39	3.42	4.26
STOI	MVDR	0.46	0.58	0.50	0.53
	Ours	0.59	0.63	0.54	0.56
PESQ	MVDR	1.72	1.95	1.86	1.71
	Ours	2.00	2.17	2.07	1.80

Table 2: Comparison of our method against MVDR for real loudspeaker experiments shown in Figure 4.

5.1 Synthetic Mixture of Speech

As there are no public datasets for audiovisual zooming, we generated data by mixing clean speech tracks in different (virtual) geometric configurations. We performed randomized trials to span the space around a given microphone array, varying the number of speakers and the solid angle over which we wished to enhance the sound. For each target solid angle, different numbers of speech signals were randomly placed within to differentiate from traditional beamforming scenarios where only one sound source is targeted.

We use the same 6-microphone hexagonal array configuration as in §4 and simulate sound source directions by delaying the audio signal at each microphone appropriately. In all cases, clean speech signals are obtained from real recordings. For each trial, we do as follows: a) starting from a selection of 10 clean speech sound tracks, we randomly choose between 2-10 overall speakers, of which 1-4 are randomly chosen to be targets and the rest are interferers; b) we randomly select a solid angle between 10° and 120° in both azimuth and elevation as well as a randomly-chosen direction-to-focus; c) given these setups, we randomly place the targets within the solid angle target-FOV as well as randomly place the interferers elsewhere. We run 500 random trials, applying both our method as well as MVDR (directed towards the center of the target FOV), and compute averaged metrics. The results are shown in Table 1.

Qualitatively, because MVDR (and other beamforming methods) only enhances sound from a single direction, when a conic angle of space contains multiple sounds, a relatively *muffled* sound enhancement results when pointing towards the FOV center. In contrast, our method integrates all sounds coming from within the desired FOV and attempts to enhance all equally, resulting in a more *crisp* sound enhancement.

5.2 Audio Speaker Experiments

We perform real experiments using four loudspeakers, playing individual sound tracks through each in various geometric configurations. The speakers are placed about a circular round-table, at 90°, 45°, 30°, and 15° (see Figure 4).

In each scenario, a single sound track is produced from each speaker and all speakers play simultaneously from different directions. We then cycle through the speakers and select either two or three adjacent speakers as the *targets*, while all others serve as the

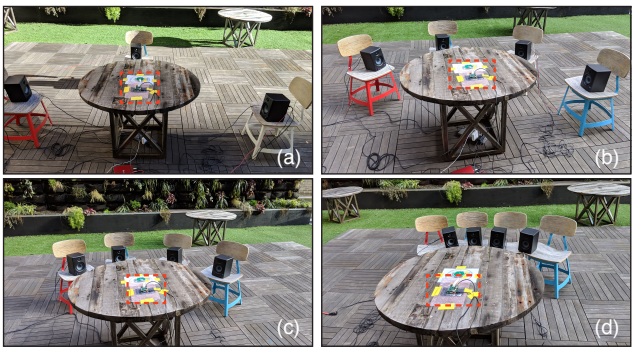

Figure 4: Loudspeaker experiments. Four audio loudspeakers play individual sounds in configurations of 90° (a), 45° (b), 30° (c), and 15° (d) angular separations. The microphone array is placed on the table (indicated by the orange boxes). We then select anywhere between 2-3 speakers to simultaneously enhance while attenuating all other speaker sounds (see supplementary video).

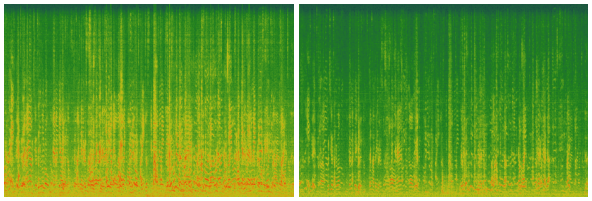

Figure 5: Speaker separation on a Ricoh camera. Using a mobile microphone array attached to a 360° camera, we perform audiovisual zooming on 4 people seated around a table at roughly 90° angular offsets from one another. In this scenario, 2 pairs of people are having simultaneous conversations and we use our method to focus in on one conversation. The left shows the raw noisy spectrogram as recorded in one of the microphones in the array. On the right, we show the spectrogram after sound enhancement using our method, which is noticeably *cleaner* (see supplementary video).

interferers. Again, we focus on more than one target sound at a time so as to differentiate from traditional beamforming scenarios. In some experiments, all speakers play clean speech tracks whereas in others, one of the speakers plays either a soft music track (e.g., jazz) or a pre-recorded "crowd noise" (e.g., recording from a crowded restaurant). Never is the music or crowd chosen as the target: these serve only to provide interference signals.

In each experiment, once the target and interferers are chosen, we play the sounds twice: a) first, all are played together to mimic a "noisy" environment; b) second, we play only the target sounds alone to serve as the "ground truth" against which we can compute SNR/SDR metrics. We present our results in each of the angular cases separately in Table 2. By all the metrics, our method outperforms MVDR.

5.3 Use Case Demonstration

Finally, we demonstrate Audiovisual Zooming on two mobile platforms: a smartphone and a 360° camera, both attached to a 6-microphone hexagonal array (see Figure 2).

We refer to our supplementary video for the audiovisual zooming demonstration using both platforms. Using the smartphone, we demonstrate a scenario in which a user captures two persons speaking simultaneously. When both persons are captured in the camera's FOV (see Figure 1-a), their voices are mixed together. As the user zooms in the camera's FOV to focus on one person (see Figure 1-b), her voice stands out while the other's voice is suppressed. Next, the user shifts the camera's FOV to another person (see Figure 1-c), and consequently his voice gradually becomes clear while the other fades out. We note that in this process the change of audio signal is fully synchronized with the change of the camera's FOV, thanks to our audiovisual zooming technique—for example, the sound gradually changes from one person's voice to another voice as the camera pans.

To demonstrate the 360° camera, four people sit around a round table and simultaneously converse. The 360° camera with the microphone array is placed on the table and pointed upwards. It is difficult to distinguish the individual speakers in the raw audio. Since the 360° camera captures all speakers, the user can set the camera's FOV on individual speakers to boost their voice relative to the others'. As the camera's FOV switches from one speaker to another, the boosted voice switches correspondingly. Consequently, the user can choose to see and hear individual speakers (see supplementary video).

In these scenarios, it was not possible to obtain ground truth (e.g., target-only sounds that perfectly match the raw, noisy signals), and so here we show our results qualitatively via spectrograms before and after enhancement (see Figure 5).

6 CONCLUSION

In this work, we extend the concept of the camera's FOV to enhance audio recording. Traditionally, camera's FOV defines only the visual frustum through which the visual content is captured by the camera. A fundamental limitation is the inconsistency between captured visual and auditory content—the sound is captured regardless of the FOV setup. To address this limitation, we have introduced an audiovisual zooming technique by leveraging the microphone array and augmenting classic beamforming methods. We have presented a method that estimates the sound spectral matrices which accounts for the desired sound signals within the FOV and those outside of the FOV. The estimated spectral matrices allow us to enhance sound coming within the FOV by solving a generalized eigenvalue problem. Our method requires no analysis of captured video frames. It can enhance however many sound sources within the FOV, and the captured imagery is in tandem with the resulting sound signal.

A limitation of our approach is that in a reflective environment, a sound source outside of the FOV may emit sound waves that arrive to the microphone from within the FOV through reflections. In this case, our audiovisual zooming method will still enhance those received sound signals. In the future, we plan to investigate this limitation by estimating the room acoustics, which might require the analysis of captured video frames to understand the environment geometry and acoustic properties (e.g., [17]).

REFERENCES

[1] 2013. *Microphone Array Beamforming*. Technical Report. InvenseSense, Inc. Document Number: AN-1140-00.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The Conversation: Deep Audio-Visual Speech Enhancement. *arXiv preprint arXiv:1804.04121* (2018).

[3] John Billingsley and R Kinns. 1976. The acoustic telescope. *Journal of Sound and Vibration* 48, 4 (1976), 485–510.

[4] Michael Brandstein and Darren Ward. 2013. *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media.

[5] Jack Capon. 1969. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* 57, 8 (1969), 1408–1418.

[6] N.Q.K. Duong, P. Berthet, S. Zabre, M. Kerdranvat, A. Ozerov, and L. Chevallier. 2017. Audio Zoom for Smartphones Based On Multiple Adaptive Beamformers. In *International Conference on Latent Variable Analysis and Signal Separation*.

[7] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman, and M. Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. In *ACM Transactions on Graphics (SIGGRAPH)*.

[8] Weijiang Feng, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo. 2017. Audio visual speech recognition with multimodal recurrent neural networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 681–688.

[9] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. 2017. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 4 (April 2017), 692–730.

[10] L.J. Griffiths and C.W. Jim. 1982. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.* 30, 1 (January 1982), 27–34.

[11] Y. Gu and A. Leshem. 2012. Robust Adaptive Beamforming Based on Interference Covariance Matrix Reconstruction and Steering Vector Estimation. *IEEE Transactions on Signal Processing* 60, 7 (July 2012), 3881–3885.

[12] John R Hershey and Michael Casey. 2002. Audio-visual sound separation via hidden Markov models. In *Advances in Neural Information Processing Systems*. 1173–1180.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach. 2016. Neural Network Based Spectral Mask Estimation For Acoustic Beamforming. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[14] Eric KL Hung and Ross M Turner. 1983. A fast beamforming algorithm for large arrays. *IEEE Trans. Aerospace Electron. Systems* 4 (1983), 598–607.

[15] Chanwoo Kim and Richard M Stern. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Ninth Annual Conference of the International Speech Communication Association*.

[16] Chiong Ching Lai, Sven Erik Nordholm, and Yee Hong Leung. 2017. *A Study Into the Design of Steerable Microphone Arrays*. Springer.

[17] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. 2018. Scene-Aware Audio for 360° Videos. *ACM Trans. Graph.* 37, 4 (2018).

[18] Jian Li, Petre Stoica, and Zhisong Wang. 2003. On robust Capon beamforming and diagonal loading. *IEEE transactions on signal processing* 51, 7 (2003), 1702–1715.

[19] Carl D Meyer. 2000. *Matrix analysis and applied linear algebra*. Vol. 71. Siam.

[20] Ulf Michel et al. 2006. History of acoustic beamforming. In *Berlin Beamforming Conference, Berlin, Germany, Nov.* 21–22.

[21] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2130–2134.

[22] Alan V Oppenheim. 1999. *Discrete-time signal processing*. Pearson Education India.

[23] Andrew Owens and Alexei A. Efros. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *ECCV*. Springer, 639–658.

[24] V. Rabinovich and N. Alexandrov. 2013. Typical Array Geometries and Basic Beam Steering Methods. In *Antenna Arrays and Automotive Applications*. Springer Science+Business Media, New York, Chapter 2, 23–54.

[25] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon Chambers. 2014. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine* 31, 3 (2014), 125–134.

[26] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, 749–752.

[27] W. Ruochen, Z. Yuhong, and Z. Wei. 2014. Acoustic Zooming Based on Real-Time Metadata Control. In *Proceedings of IC-NIDC*.

[28] Petre Stoica, Randolph L Moses, et al. 2005. Spectral analysis of signals. (2005).

[29] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4214–4217.

[30] Heinz Teutsch. 2007. *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Vol. 348. Springer.

[31] O. Thiergart, K. Kowalczyk, and E.A.P. Habets. 2014. An Acoustical Zoom Based on Informed Spatial Filtering. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*.

[32] H.L. Van Trees. 2002. *Detection, Estimation, and Modulation Theory, Part IV: Optimum Array Processing*. Wiley, New York.

[33] B.D. Van Veen and K.M. Buckley. 1988. Beamforming: A versatile approach to spatial filtering. *IEEE Acoust., Speech, Signal Process. Mag.* 5, 2 (April 1988), 4–24.

[34] E. Vincent, R. Gribonval, and C. Fevotte. 2006. Performance Measurement in Blind Audio Source Separation. *Transactions on Audio, Speech and Language Processing* 14, 4 (July 2006), 1462–1469.

[35] E. Warsitz and R. Haeb-Umbach. 2007. Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 5 (July 2007), 1529–1539.

[36] Jung-Lang Yu and Chien-Chung Yeh. 1995. Generalized eigenspace-based beamformers. *IEEE Transactions on Signal Processing* 43, 11 (1995), 2453–2461.

[37] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The Sound of Pixels. In *The European Conference on Computer Vision (ECCV)*.

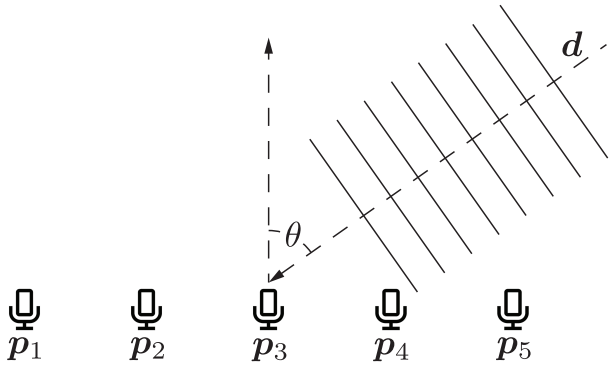

Figure 6: Consider a microphone array in which each microphone is located at position p_i. A single sound comes from the direction d as a plane wave. The angle between the microphone array's facing direction and the sound incoming direction is θ.

A SPECTRAL MATRIX OF SOUND FROM A DIRECTION θ

Consider a plane wave impinging on a microphone array (see Figure 6). Let $\boldsymbol{p}_i$ denote the position of individual microphones, and the plane wave comes from the direction $\boldsymbol{d}$ with an intensity A. The angle between the sound incoming direction and the microphone array's facing direction is θ. Then, the sound waves received at individual microphone is expressed as

$$s_i = A^{\frac{1}{2}} e^{-j(\frac{\omega}{c}\boldsymbol{d}^T\boldsymbol{p}_i + \omega t)}.$$

Here $\frac{\omega}{c}\boldsymbol{d}^T\boldsymbol{p}_i$ is the (relative) phase at the microphone i. Putting all s_i into a vector $\boldsymbol{s} = [s_1 \ \dots \ s_M]^T$, we can compute the spectral matrix by its definition [22] as

$$\mathrm{R}(\omega) = \mathrm{FFT}\{\boldsymbol{s}\boldsymbol{s}^T\} = A\boldsymbol{v}_{\boldsymbol{d}}\boldsymbol{v}_{\boldsymbol{d}}^H, \tag{16}$$

where $\boldsymbol{v}_{\boldsymbol{d}}$ is the steering vector defined in (6). This expression (16) is what we used in (11).

B DERIVATION OF ERROR BOUND (15)

First, we substitute (13) into the $\mathrm{R_s}$ estimation (11) and obtain the expression of Δ in (14),

$$\Delta = \int_{\boldsymbol{\Theta}} \frac{1}{a}\boldsymbol{v}_\theta\boldsymbol{v}_\theta^H = \int_{\boldsymbol{\Theta}} \frac{1}{\boldsymbol{v}_\theta^H \mathrm{R_c}^{-1}\boldsymbol{v}_\theta}\boldsymbol{v}_\theta\boldsymbol{v}_\theta^H. \tag{17}$$

Here $\boldsymbol{v}_\theta^H \mathrm{R_c}^{-1}\boldsymbol{v}_\theta$ is bounded by the maximum and minimum eigenvalue of $\mathrm{R_c}^{-1}$. Also, notice that $\boldsymbol{v}_\theta$ is the steering vector, which has a specific form (6). Therefore, we have

$$\hat{\lambda^c_{\min}}\boldsymbol{v}_\theta^H\boldsymbol{v}_\theta^H = \hat{\lambda^c_{\min}}M \le \boldsymbol{v}_\theta^H\mathrm{R_c}^{-1}\boldsymbol{v}_\theta \le \hat{\lambda^c_{\max}}\boldsymbol{v}_\theta^H\boldsymbol{v}_\theta^H = \hat{\lambda^c_{\max}}M, \tag{18}$$

where M is the number of microphones; $\hat{\lambda^c_{\max}}$ and $\hat{\lambda^c_{\min}}$ are the maximum and minimum eigenvalues of $\mathrm{R_c}^{-1}$, respectively. They are related to the eigenvalues of $\mathrm{R_c}$ through

$$\hat{\lambda^c_{\max}} = \frac{1}{\lambda_{\min}} \text{ and } \hat{\lambda^c_{\min}} = \frac{1}{\lambda_{\max}}.$$

Combing this expression with (17) and (18), we obtain the error bound of Δ as shown in (15).

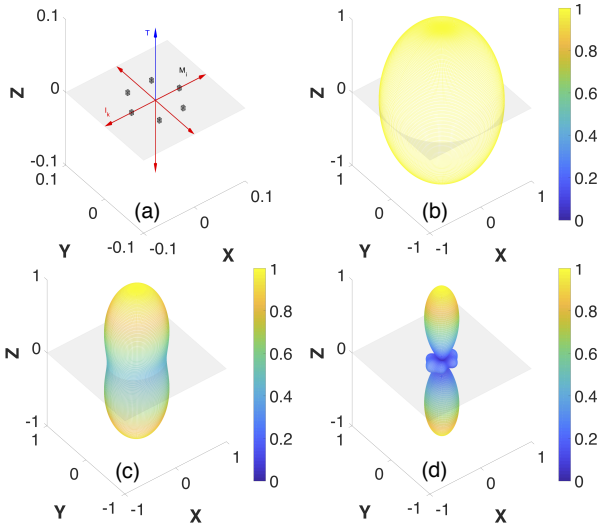

Figure 7: Frequency dependence. We visualize the frequency dependence of the MVDR beam patterns. (a) The array consists of 6 microphones shown as gray cubes on the X-Y plane, where the microphones are spaced evenly 5cm from one another. The 6 sound sources are spread throughout the space: 4 interfering sources are shown in red on the X-Y plane along with another interfering source in the negative z-axis. The target is shown in blue in the positive z-axis. (b-d) The MVDR beam patterns at three different frequencies, 300Hz (b), 1860Hz (c), and 3420Hz (d), are shown both in the shape of the surface and as the color (yellow as 1 and blue as 0).

C DETAILS OF EMPIRICAL STUDIES

In our empirical studies, the MVDR beamformer aims to enhance the sound coming from the positive Z-direction, while suppressing everything else. Since we know precisely what the unwanted signals are in our simulation, we can directly compute the noise spectral matrix $\mathrm{R_n}$, which is in turn used in (7) to evaluate $\boldsymbol{w}_{\mathrm{BF}}$. We visualize MVDR performance by evaluating its *beam pattern* across a range of frequencies. The beam pattern describes the effective gain for signals coming from individual directions $\boldsymbol{\theta}$ when the beamformer is set to enhance toward a direction $\boldsymbol{\theta}_0$. It is defined as

$$g(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = |\boldsymbol{w}^H_{\mathrm{BF},\theta_0}\boldsymbol{v}_\theta|, \tag{19}$$

where $\boldsymbol{w}^H_{\mathrm{BF},\theta_0}$ is the MVDR weights from (7) when the steering direction is set to be $\boldsymbol{\theta}_0$, and $\boldsymbol{v}_\theta$ is defined in (6).

Frequency dependence. To study the frequency dependence of the microphone array's performance, we use sound sources that produce sound signals at a fixed frequency, and the frequency is varied to ascertain the beamforming performance with respect to frequency change. This is seen in Figure 7 as tighter main lobes in the target's direction for higher frequencies, meaning the interfering sounds are better suppressed.

Array size. Figure 8 shows the change in *average beam pattern* across the human speech frequency range (300-3420Hz) as the microphone spacing is changed from 0.5cm to 5cm and then to 50cm. The more microphones we have, the better we can sample

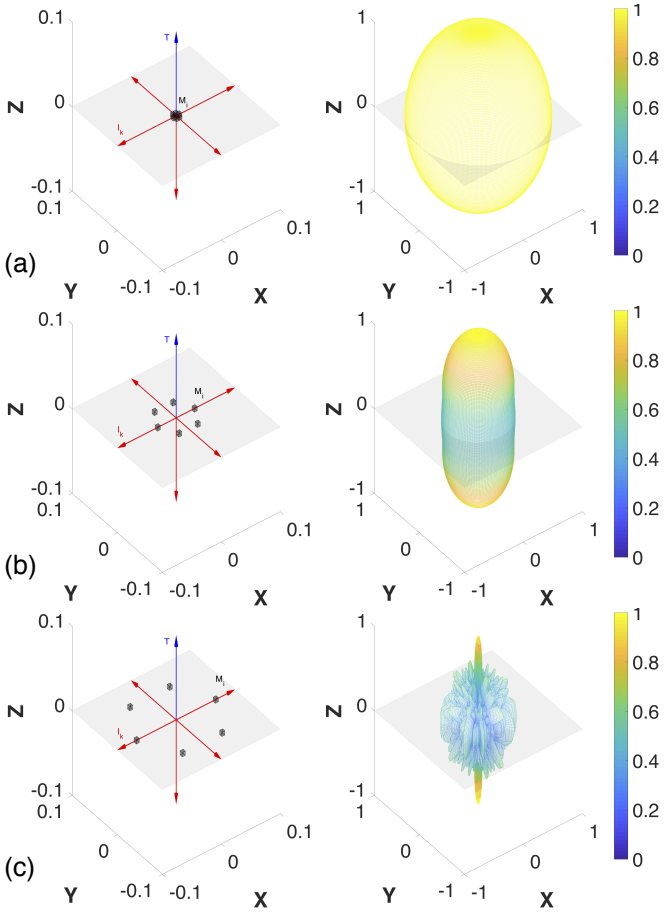

Figure 8: Role of array size. We use a 6-element circular array but vary the inter-microphone spacing to adjust the overall array size. On the left of each subplot is the spatial configuration: the target, interferers, and ambient noise are the same as before (Figure 7), and the spacing of the microphones (in the x-y plane) changes from 0.5cm (a) to 5cm (b) to 50cm (c). On the right of each subplot is the average beam pattern across the frequency range of human speech, to indicate the average performance of the beamformer in that range.

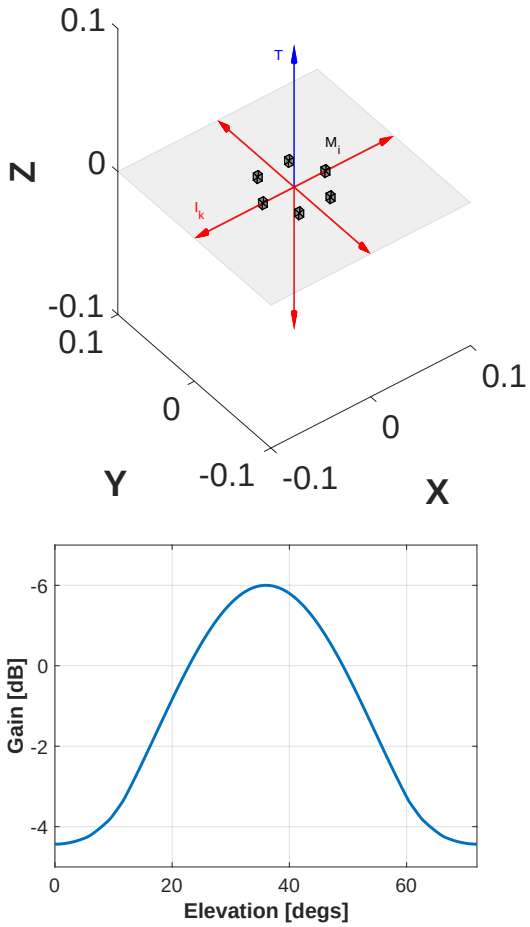

Figure 9: Target proximity sensitivity. When beamforming at a target in the presence of interferers, it is important to know how the gain falls off from the direction of the target for nearby interferers. We show this for a given 6-microphone array configuration with 5cm spacing [top] as a 3D surface plot of average gain (across the human frequency range) as a function of azimuth and elevation angular offset from the desired target direction [bottom]. This allows us to better understand how close sounds can be before they are not sufficiently separable.

(spatially) with larger array sizes. The smallest spacing setting of 0.5cm gives almost no directionality, with an omni-directional gain response. As we increase to 5cm, the directionality improves with better suppression of the interference sources relative to the target.

Number of microphones. The simulation setup and results are shown in Figure 10 and its caption.

Sampling density. Figure 9 (bottom) shows a plot of average gain (y-axis) within the human frequency range as a function of elevation (x-axis) angle (e.g., offset from the target direction). Note that we ignore the variation of azimuth angle because it has no effect for the given array configuration and target direction, as shown in Figure 7. Therefore, we consider the microphone array scenario as shown on the top (circular 6-sensor array with 5cm spacing). The simulation shows that the gain falls off from the target direction for any nearby sounds within a small FOV of the target. We use this to determine a reasonable sampling rate for our sphere integration approach, as discussed in the main text (in §4). Here, we convert the gain as expressed in (19) to dB as: $g_{dB}(\theta) = 20 * log_{10}(g(\theta))$.

Extension to 3D arrays. We explore the effect of a 3D microphone array as a future extension. 3D array is able to break the symmetry that a 2D array suffers from, although it is much more bulky and might not be compatible with the small form factor of most mobile devices. The result and simulation details are shown in Figure 11 and its caption.

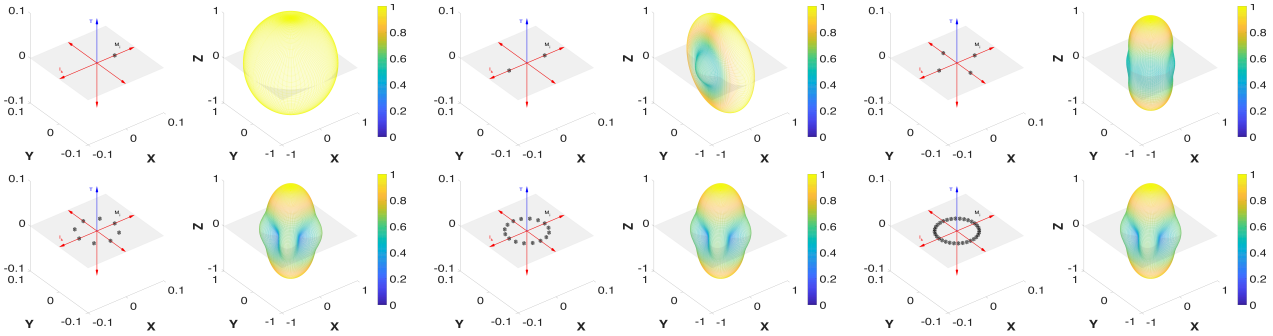

Figure 10: Here the microphone array geometry is a circle with a fixed 5cm radius in the X-Y plane. We examine how changing the number of microphones on this circule affects the average beam pattern of the beamformer. The definition of beam pattern is presented in Appendix C. [Top-Left] A single microphone yields an omnidirectional response. [Top-Middle] Two microphones improves directionality by suppressing two side interferers, but not the others. [Top-Right] Four microphones improves directionality further. [Bottom-Left] Eight microphones are better, and the performance plateaus as 16 [Bottom-Middle] or 32 [Bottom-Right] microphones yield no clear improvement.

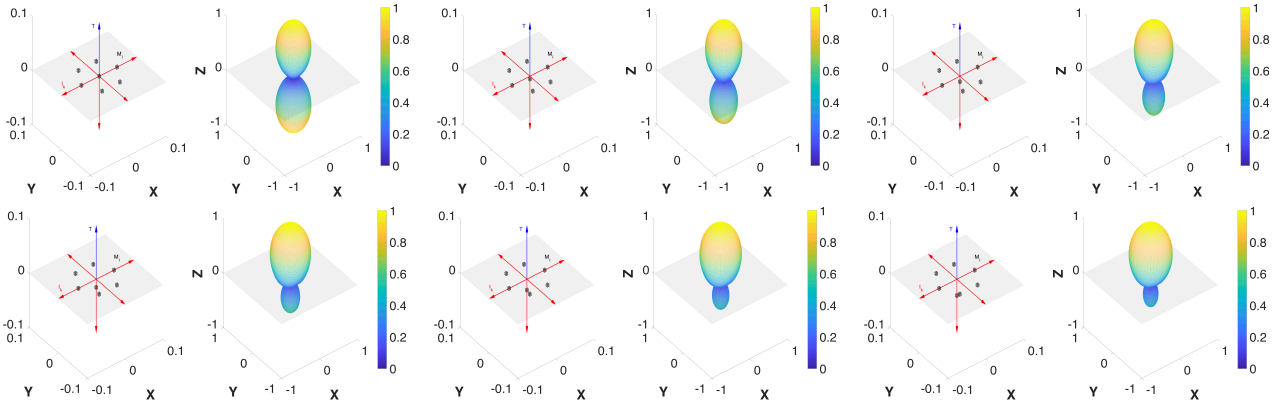

Figure 11: 3D asymmetries. Effect of MVDR beamforming as a microphone as added to the third dimension. We consider our baseline 6-microphone circular planar array and we add an extra microphone at the center. We then move the extra mic along the negative z-dimension to break the 2D symmetry and observe how this affects the gain for the previously-ambiguous interference behind the array. [Top-Left] The extra mic is at z=0, and so the symmetry remains. [Top-Middle] The extra mic moves down the negative z-axis by 5mm, and the gain in the direction of the interferer subsides. As the microphone moves further along the axis by 10mm [top-right], 15mm [Bottom-Left], 20mm [Bottom-Middle], and 30mm [Bottom-Right], the gain in the direction of the interferer attenuates more and more while the gain in the direction of the target remains maximal.