Perspectives on microphone array processing including sparse recovery, ray space analysis, and neural networks

Craig T. Jin\textsuperscript{1,*}, Shiduo Yu\textsuperscript{1,†}, Fabio Antonacci\textsuperscript{2,‡} and Augusto Sarti\textsuperscript{2,§}

\textsuperscript{1}School of Electrical and Information Engineering, University of Sydney, Sydney, Australia
\textsuperscript{2}Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Abstract: Hands-free audio services supporting speech communication are playing an increasingly ubiquitous and foundational role in everyday life as services for the home and work become more automated, interactive and robotic. People will speak their instructions (e.g. Siri) to control and interact with their environment. This makes it an exciting time for acoustics engineering because the demands on microphone array performance are rapidly increasing. The microphone arrays are expected to work at increasing distances in noisy and reverberant situations; they are expected to record not just the sound content, but also the sound field; they are expected to work in multi-talker situations and even on moving, robotic platforms. Audio technology is currently undergoing rapid change in which it is becoming feasible, from both a cost and hardware point-of-view, to incorporate multiple and distributed microphone arrays with hundreds or even thousands of microphones within a built environment. In this review paper, we consider microphone array signal processing from two relatively recent vantage points: sparse recovery and ray space analysis. To a lesser extent, we also consider neural networks. We present the principles underlying each method. We consider the advantages and disadvantages of the approaches and also present possible methods to integrate these techniques.

Keywords: Microphone array, Sparse recovery, Ray space, Convolutional neural networks

PACS number: 43.60.Fg, 43.60.Jn, 43.60.Np  [doi:10.1250/ast.41.308]

1. INTRODUCTION

We consider microphone arrays and their support for human-robot interaction in noisy, multi-talker, and reverberant environments. With microphone arrays, we assume we can incorporate multiple and distributed microphone arrays with hundreds or even thousands of microphones in a built environment. Our focus is how do we use the microphone arrays to best advantage. We review the relatively new methods of sparse recovery, ray space analysis and neural networks. Microphone arrays must perform many services such as sound source tracking, direction of arrival estimation, source separation, speech recognition, and emotion recognition. In this review, we focus on the simplest of problems — the analysis or decomposition of a sound field into plane-wave source signals and their directions. The history of microphone array processing is replete with numerous methods for sound field analysis. For example, there are simple beamforming methods that can provide an energy map of the sound field [1] and [2]. There are numerous time difference of arrival (TDOA) methods based on the cross-correlation of microphone signals (for a detailed review refer to [3]). There are also more elaborate techniques such as multiple signal classification (MUSIC) [4] and the estimating signal parameters via rotational invariance technique (ESPRIT) [5] which have recently been applied to spherical microphone arrays (see EB-MUSIC [6] and EB-ESPRIT [7]). There are also direction-of-arrival techniques based on multi-dimensional maximum-likelihood estimation which have recently been proposed for spherical microphone arrays [8]. In this review, we describe a sparse recovery method for sound field decomposition that provides increased spatial resolution compared to other methods; a ray space analysis technique that enables the integration of multiple viewpoints from a single, consistent geometric framework; and to a much lesser extent the upcoming neural network approaches for robust microphone array processing.

1.1. Source Decomposition Problem

Consider the observation of a sound field via a set of $K$
microphone signals which have had a short-time Fourier transform applied, so that signals are expressed in the time-frequency domain. We represent the microphone signals using the notation of a three-dimensional array: \( \mathcal{X}[k,t,f] \), with different microphones, indexed by \( k \), running along the first dimension; time, \( t \), running along the second dimension; and frequency, \( f \), running along the third dimension. Because analysis is performed for each frequency separately, we consider a slice of the microphone array signals for a fixed, given frequency \( f \). The slice is represented by the observation matrix \( X(f) \):

\[
X(f) = \mathcal{X}[\cdot,\cdot,f],
\]

where the matrix element \( X_{nk}(f) \) is given by \( \mathcal{X}[k,t,f] \). We also represent source signals using the notation of a three-dimensional array: \( \mathcal{S}[n,t,f] \), with different sources, indexed by \( n \), running along the first dimension; time, \( t \), running along the second dimension; and frequency, \( f \), running along the third dimension. For a fixed frequency \( f \), a slice of the source signals is represented by the source matrix \( S(f) = \mathcal{S}[\cdot,\cdot,f] \).

We assume that the relationship between the observation matrix, \( X(f) \), and the source matrix, \( S(f) \), is known and specified by the transformation matrix, \( A(f) \):

\[
A(f)S(f) = X(f).
\]

The transformation matrix, \( A(f) \), is the matrix expressing the contribution of the different source signals to the observation signals:

\[
A(f) = [a(\Omega_1,f), a(\Omega_2,f), \ldots, a(\Omega_N,f)]
\]

\[
a(\Omega_n,f) = [a_1(\Omega_n,f), a_2(\Omega_n,f), \ldots, a_K(\Omega_n,f)]^T,
\]

where \( a_k(\Omega_n,f) \) represents the complex-valued gain between the source signal, with frequency \( f \), incoming from location \( \Omega_n \) and the \( k \)-th observation signal. The vector \( a(\Omega_n,f) \) is commonly referred to as an array manifold vector.

Using the above formalism, we describe the source decomposition problem as solving Eq. (2) for the unknown source matrix, \( S(f) \), given both the observation matrix, \( X(f) \), and the transformation matrix, \( A(f) \). In other words, we want to determine the source signals given the observation signals and knowledge of the array manifold vectors. One difficulty is that the set of possible source signal directions is infinite. In order to resolve the source direction with high-resolution, one includes a large number of array manifold vectors into the matrix, \( A(f) \), one for each direction considered. Each possible direction considered corresponds to a column of matrix \( A(f) \) with the implication that \( A(f) \) becomes a very wide matrix leading to an under-determined problem. In the ensuing discussion, we refer to \( A(f) \) as a dictionary matrix—a source direction dictionary—because we are expressing the observation signals as a sum of source contributions that are defined by \( A(f) \).

## 2. SPARSE SOURCE DECOMPOSITION

The trade-off between increasing the resolution of the source dictionary and having an increasingly under-determined problem leads to consideration of sparse recovery methods. To begin, assume that the rank of the dictionary matrix \( A \) is \( K \), the number of microphone signals. This will generally be true when the directions in the dictionary matrix are distributed sufficiently evenly across space. The source decomposition problem is then under-determined because there is an infinite number of valid and possible solutions to Eq. (2). A classic way to circumvent this problem is to choose the solution, \( S(f) \), with the least energy, that is, the smallest Frobenius norm (or \( \ell_2 \) norm when there is only one time sample). This solution is known as the least-norm solution and is given by:

\[
\hat{S}(f) = A^H(f)(A(f)A^H(f))^{-1}X(f).
\]

The matrix \( A^H(f)(A(f)A^H(f))^{-1} \) is often referred as the Moore-Penrose pseudo-inverse of \( A \). The trouble with the least-norm solution is that it tends to distribute the energy evenly across source directions. This physical assumption is generally wrong and leads to an undesirable spatial blurring of the acoustic image.

A better alternative to the least-norm solution is to consider the sparsest solution, that is, the solution with the fewest number of plane waves that still explains the observations. Mathematically, this solution can be defined as:

\[
\hat{S}(f) = \text{argmin}_{S(f)} ||S(f)||_{0.2} : A(f)S(f) = X(f),
\]

where \( || \cdot ||_{0.2} \) denotes the \( \ell_{0.2} \) norm. The \( \ell_{0.2} \) norm of matrix \( S \) is defined as the total number of source signals for which the corresponding signal energy is non-zero, i.e., the total number of indices \( n \) for which:

\[
\sum_{t=1}^{T} |S_{nt}(f)|^2 \neq 0.
\]

This solution is advantageous for a few reasons. First, it provides a sharper acoustic image than the least-norm solution because it finds the fewest number of plane waves that explain the observations, i.e., the solution is spatially sparse. As well, it is reasonable to apply the principle of Occam’s razor and explain the observations in terms of the fewest number of simultaneous sources.

### 2.1. The Iteratively-reweighted Least-square Algorithm

In practice it is extremely difficult to solve Eq. (5) for
Mathematically, this also promotes sparsity. The least Eq. (9) depend on matrix method of the Lagrange multipliers. Because the weights in Note that this result can be easily demonstrated using the corresponding to the different time instants are not linearly independent and the observation signals can be replaced by their PCA representation, which is a $K \times K$ matrix. For further details on the use of PCA to accelerate the computations, please refer to [14].

The IRLS algorithm is shown in Fig. 1 and presented here for completeness. Note that the parameter $\beta$ representing the relative energy of the noise can be estimated as described in [14,15]. The weights are first all initialized to one. The algorithm then repeats the following two steps in alternation until convergence: 1) update the solution $S$ assuming a fixed weight matrix $W$ (line 7 of the algorithm); and 2) update the weight matrix given the solution

\[
\begin{align*}
1: & \text{ Inputs: plane-wave dictionary } A, \text{ observation signals } B, \text{ relative energy of noise } \beta \\
2: & \text{ Outputs: plane-wave signals } X \\
3: & \text{ Initialization:} \\
4: & W = I_N \\
5: & \text{ Until convergence, do:} \\
6: & \gamma \leftarrow \frac{\beta}{N} \frac{1}{1-\beta} \text{tr}(AWA^H) \\
7: & X \leftarrow W A^H (AWA^H + \gamma I)^{-1} B \\
8: & \text{ for } n = \{1, 2, \ldots, N\}, e_i \leftarrow \sum_{t=1}^{T} |X_{n,t}|^2 \\
9: & \epsilon_{\text{max}} \leftarrow \max\{e_i, i = 1, \ldots, N\} \\
10: & \epsilon \leftarrow \min\left(\epsilon, \frac{\epsilon_{\text{max}}}{\rho}\right) \\
11: & \text{ for } i = \{1, 2, \ldots, N\}, w_i \leftarrow \rho_i^{-\beta} (e_i + \epsilon)^{-\frac{1}{2}} \\
12: & W \leftarrow \text{diag}(w_1, w_2, \ldots, w_N)
\end{align*}
\]

Fig. 1 The regularized IRLS algorithm for sparse source decomposition. The algorithm includes a prior expectation, $\rho$, imposed on the source directions determined, e.g., by an MPDR beamformer.

\[
\begin{align*}
\text{The idea behind the IRLS algorithm is that the } \ell_{p,2} \text{ norm of the solution can be expressed as a weighted } \ell_{2,2} \text{ norm (the Frobenius norm). We have:} \\
\|S(f)\|_{\ell_{p,2}}^p & = \sum_{n=1}^{N} \left( \sum_{t=1}^{T} |S_{n,t}(f)|^2 \right)^{\frac{p}{2}} \\
& = \sum_{n=1}^{N} \left( \sum_{t=1}^{T} |S_{n,t}(f)|^2 \right)^{\frac{p-2}{2}} \left( \sum_{t=1}^{T} |S_{n,t}(f)|^2 \right)^{\frac{p}{2}} \\
& = \|W^{-\frac{1}{2}}(f)S(f)\|_2^2, \\
\text{where } W(f) \text{ is the diagonal matrix given by:} \\
W(f) & = \text{diag}(w_1(f), w_2(f), \ldots, w_N(f)) \\
w_n(f) & = \left( \sum_{t=1}^{T} |S_{n,t}(f)|^2 \right)^{\frac{2-p}{2}}. \\
\end{align*}
\]

Therefore, finding the solution with the least $\ell_{p,2}$ norm is equivalent to solving a weighted least-norm problem in which the weights depend on the solution. So long as there is little danger of confusion, in the ensuing discussion we omit the frequency dependence of the variables for the purpose of notational simplicity.

The IRLS algorithm solves the following weighted least-norm problem:

\[
\begin{align*}
\text{minimize } & \|W^{-\frac{1}{2}}S\|_2^2 \text{ subject to } AS = X. \\
\text{For fixed } W, \text{ this problem has a closed-form solution, } S_W, \text{ given by:} \\
S_W & = W A^H (AWA^H)^{-1} X. \\
\end{align*}
\]

Note that this result can be easily demonstrated using the method of the Lagrange multipliers. Because the weights in Eq. (9) depend on matrix $X$, the IRLS algorithm calculates the solution iteratively as summarized in Fig. 1.

2.2. Practical Sparse Source Decomposition

The accuracy and robustness of sparse recovery methods improve when prior information regarding the support of the solution is available [11,12]. In practice, we can incorporate prior information into the sparse recovery using a weighted $\ell_{p,2}$-norm. In other words, we solve:

\[
\begin{align*}
\tilde{S}(f) & = \arg\min_{S(f)} \|\Pi(f)S(f)\|_{\ell_{p,2}} : A(f)S(f) = X(f), \\
\text{where } \Pi(f) \text{ is a diagonal matrix weighting the possible source signals. The diagonal elements, } \rho(\Omega_n, f) \leq 1, \text{ of the weighting matrix } \Pi \text{ are commonly set according to the inverse of the minimum power distortionless response (MPDR) beamforming energy:} \\
\rho(\Omega_n, f) & = a^\top(\Omega_n, f)R^{-1}(f)a(\Omega_n, f), \\
\text{where } R(f) = E[X(f)X(f)^\top] \text{ is the covariance matrix of the microphone signals and } E[\cdot] \text{ is the expectation operator.}
\end{align*}
\]

It is also important to note that when the number of time samples $T$ is larger than the number of observation signals, $K$, principal component analysis (PCA) may be used to accelerate the computations, as proposed in [13]. Indeed, if $T > K$, the vectors of observation signals corresponding to the different time instants are not linearly independent and the observation signals can be replaced by their PCA representation, which is a $K \times K$ matrix. For further details on the use of PCA to accelerate the computations, please refer to [14].
2.3. Sparse Decomposition Performance

The performance of the sparse decomposition method is demonstrated experimentally. The results that follow are taken from [14]. We apply the sparse source decomposition algorithm to the analysis of signals recorded in a office room with a spherical microphone array (SMA). A diagram illustrating the measurement setup is shown in Fig. 2. The microphone array was the dual-concentric SMA described in [16]. This SMA consists of 64 microphones located on the surface of two spheres: 32 microphones are distributed evenly on a rigid sphere of radius 28 mm and the 32 others are distributed on an open sphere of radius 95.2 mm. Impulse responses were measured using an NTI-Audio Talkbox [17] loudspeaker which was moved to the three positions. As well, the loudspeaker pointed directly at loudspeaker directions $S_1$ and $S_2$, located at azimuths $0^\circ$ and $45^\circ$, respectively, are not resolved by the spherical beamformer. Compared to the spherical beamforming map, the map obtained using the MUSIC algorithm has a higher resolution and presents a clear peak in each of the three speaker directions. However, the amplitude of these peaks does not reflect the power of the sources accurately. Given that the speech signals have equal powers, and given the distance at which the speakers are located, the power received at the SMA from source $S_1$ (located at azimuth $-135^\circ$) should be approximately 3.5 dB lower than that received from sources $S_1$ and $S_2$. By contrast, the peak observed in the MUSIC spectrum in the direction of $S_3$ is about 12 dB lower than that observed in the directions of $S_1$ and $S_2$, which is understandable given that the MUSIC spectrum is not designed to correctly indicate source power.

Still referring to Fig. 3, the map obtained using the sparse recovery method presents five sharp energy peaks. The three highest peaks are within a few degrees of the loudspeaker’s membrane center matched that of the SMA center, 1.5 m, therefore the elevation of the speaker relative to the SMA was $0^\circ$ for the three positions. As well, the loudspeaker pointed directly at the SMA in each location. The room in which the impulse responses were measured is a office space with dimensions $14 \times 8 \times 3$ m, approximately. The room’s measured T60 reverberation time is about 0.5 s and the direct-to-reverberant ratio (DRR) was estimated to be 7 dB when the speaker was located 1.8 m from the SMA and 4.5 dB when the speaker was located 2.7 m from the SMA. The impulse responses measured with the SMA for the three speaker locations were convolved with three anechoic speech signals to generate echoic microphone signals.

We analyzed the spherical harmonic signals using three different spatial sound field analysis techniques. First, we applied spherical beamforming and steered the beam in a large number of directions over the sphere. We then calculated the power of the beamformer output for every steering direction to obtain a map of the incoming sound energy. Second, we applied the MUSIC algorithm [4] to the spherical harmonic signals. The MUSIC spectrum for direction $\Omega_n$ is given by:

$$
\mu(\Omega_n) = \left( \sum_{i=4}^{K} a(\Omega_n)^T u_i \right)^{-1}
$$

where $u_i$ is the $i$-th eigenvector of the spherical harmonic signal covariance matrix. Note that the calculation of the MUSIC spectrum requires knowledge of the number of sources. In the formula above, only the eigenvectors $u_i$ corresponding to the noise subspace are summed. The summation starts with $i = 4$ because there are three sound sources. Lastly, we performed a sparse plane-wave decomposition of the band-passed spherical harmonic signals by decomposing the signals over a dictionary of 3,074 plane-wave directions using the regularized IRLS algorithm. We then calculated the power of the plane-wave signals, thus obtaining a map of the incoming acoustic energy.

The results are shown in Fig. 3. Although the map resulting from the spherical beamforming technique presents higher power values in the directions where the speakers are located, its resolution is very low. In particular sources $S_1$ and $S_2$, located at azimuths $0^\circ$ and $45^\circ$, respectively, are not resolved by the spherical beamformer.
power of the sources, the peak corresponding to $S_1$ is about 6 dB lower than the peaks observed in the directions $S_1$ and $S_2$, which is relatively close to the expected 3.5 dB difference. In summary, the sparse recovery algorithm made it possible to detect the three speaker locations and power with excellent accuracy, as well as two of the reflections occurring on the wall. This work is supported by numerous other similar demonstrations [13,18–20].

3. RAY SPACE ANALYSIS

In this section, we review Ray Space Analysis (RSA) [21] and describe both the Euclidean ray space and projective ray space. We show that an advantage of RSA is the unified and consistent geometric framework that it provides when working with multiple microphone arrays. Given the mathematical complexities of RSA, we consider only linear microphone arrays in this review. A core idea underlying RSA is the consideration that a long linear microphone array is composed of many sub-arrays. Each sub-array has its own directional characteristics and provides a specific view of the sound field. One can combine the acoustic image from each sub-array to obtain an integrated picture that is referred to as the Euclidean ray space. In the Euclidean ray space, a straight line represents a point source in geometry space. This will be detailed below shortly. When one has multiple microphone arrays, one obtains a sound field description composed of multiple Euclidean ray spaces. It is difficult to integrate geometrical information across multiple Euclidean ray spaces. For this reason, the projective ray space has been proposed [22] to unite and integrate geometrical information across multiple Euclidean ray spaces onto a single sphere.

The Euclidean ray space provides a means to describe the sound field captured by an array with $K$ microphones, where a sound ray is expressed as one straight line: $y = mx + q$, foregoing rays parallel to the $y$-axis, where $m = \tan(\theta)$ is the slope, $q$ is the intercept of this line and $y$-axis, and $\theta$ is the inclination of the slope. Consider now Fig. 4. The linear microphone array is shown to the left with eight microphones. The coordinate axes have been chosen such that the microphone array is along the $y$-axis. This special choice for the coordinate axes makes the interpretation of the Euclidean ray space more intuitive. The microphone array has been divided into six subarrays. For presenting the sound field information received by each viewpoint, the array is partitioned into $W = K - L + 1$ sub-arrays, each formed by $L$ microphones. In this particular example, $K = 8, L = 3$, so that $W = 8 - 3 + 1 = 6$. The ray slopes, $m$, are divided into $M$ slope values. Using the time-frequency notation specified earlier, the array data for the microphones in the $w$-th sub-array is then given by: $\mathcal{X}[w \cdots w + L - 1, t, f]$. The computation of the ray space transform, $\mathcal{R}(m, q, f)$, proceeds through the application of a spatial window, $\Psi$, to the given sub-array, $\mathcal{X}[w(q) \cdots w(q) + L - 1, t, f]$ followed by the computation of a delay-and-sum beamformer to yield:

$$\mathcal{R}(m, q, f) = \sum_{l=0}^{L-1} \mathcal{X}[w(q) + l, t, f] \exp\left(-j k \frac{y(w(q) + l)m}{\sqrt{1 + m^2}}\right) \Psi(l),$$

where $w(q)$ is the microphone index for ray position $q$, $y(k)$ is the $y$-coordinate of the $k$-th microphone, $k = \omega/c$ is the wavenumber, and $\Psi(l)$, $l = 0, \ldots, L - 1$ is the adopted spatial window function.
In the special case where the $y$-axis lies along the microphone array, $q$ simply indicates the position of a given subarray along the $y$-axis. Furthermore, the slope, $m$, indicates the beamforming direction for the subarray. Thus, the beamforming energy in various directions for a given subarray forms one horizontal slice of the Euclidean ray space image as shown at the bottom of Fig. 4. The acoustic source is then represented as a line (the bright region in the image) in the Euclidean ray space image. It is important to note that when the $y$-axis does not lie along the microphone array, then the value of $q$ can vary with $m$. Nevertheless, a single acoustic source is always represented by a line in the Euclidean ray space image.

The difficulty with multiple microphone arrays is the lack of a common geometric basis from which to interpret the multiple Euclidean ray space images. For example, in Fig. 5 two different linear arrays are shown perpendicular to each other. A single acoustic source results in a line in the Euclidean ray space image corresponding to each linear array. However, there is no clear indication that the two lines correspond to a single source.

In [22], the projective ray space was introduced to provide a single, unifying geometric framework from which to analyze multiple linear microphone arrays. In a sense, what one does is to project multiple Euclidean ray space images onto a single sphere. An acoustic source is then identified by a great circle on that sphere. To clarify these issues, first consider Fig. 6. One chooses a single Cartesian axes for all of the different microphone arrays.

For the top row of Fig. 6, the Cartesian axes is oriented so that the $y$-axis lies along the vertical microphone array. In the bottom row, the Cartesian axes is placed so that the origin is located at the position of the acoustic source (marked by a cross). When there are multiple microphone arrays, the interpretation and intuition becomes much clearer when the origin of the Cartesian axes is placed at the position of the acoustic source. In this case, $q = -x_0m + y_0 = 0$ because the source position, $(x_0, y_0)$, is the origin, $(0, 0)$. In other words, the acoustic source corresponds to the $m$-axis. However, not all of the $m$-axis is exposed by the two microphone arrays. One must actually completely surround the acoustic source by microphone arrays in order to expose the entire $m$-axis. The $m$-axis is related to the slope or the angle with which one views the acoustic source.
and the entire $m$-axis will only be exposed when one views the acoustic source from all angles.

With these considerations in mind, it becomes easier to explain the projective ray space. Consider now Fig. 7. The first issue is defining homogeneous coordinates to realize a projective space. To establish a correspondence between the two-dimensional Euclidean ray space spanned by $(m, q)$ and the three-dimensional projective ray space spanned by the vector $l = [l_1, l_2, l_3]^T$, we use (note that $m = \tan(\theta)$):

$$\begin{align*}
l_1 &= \frac{\sin(\theta)}{\sqrt{1 + q^2 \cos^2(\theta)}} \\
l_2 &= -\frac{\cos(\theta)}{\sqrt{1 + q^2 \cos^2(\theta)}} \\
l_3 &= \frac{q \cos(\theta)}{\sqrt{1 + q^2 \cos^2(\theta)}}.
\end{align*}$$

Note the vector, $l$, has unit length. In this way, we can express $R(m, q, f)$ as $R(-l_1/l_2, -l_3/l_2, f)$. In Fig. 7(a), we have surrounded the acoustic source by a four linear microphone arrays. Therefore, we expose the entire $m$-axis. Using the projective mapping, the $m$-axis becomes a great circle on the unit sphere (Fig. 7(b)). The Euclidean ray space image for each microphone array becomes a sector of the unit sphere as shown in Fig. 7(c).

In practice, one may not have the same Cartesian axes for all microphone arrays. In this case, once a given projective ray space has been chosen, one can relate all other projective ray spaces to the chosen projective ray space via a geometrical transformation for the $i$-th array (based on rotation and translation and defined by $H^{(i)}$):

$$l = (H^{(i)})^{-T}l^{(i)}. $$

The transformation matrix, $H^{(i)}$, is defined as follows:

$$H^{(i)} = \begin{bmatrix} R^{(i)} & t^{(i)} \\ 0 & 1 \end{bmatrix}, $$

where $R^{(i)}$ is a rotation matrix and $t^{(i)}$ is a translation vector.

Given the complexities of the ray space description, one may not yet clearly see the full benefits. One benefit is surely the unified geometrical framework, however, this is not all. Another benefit is the ability to resolve ambiguities that arise from the multiple viewpoints provided by the multiple microphone arrays. To understand this point, consider Fig. 8. Each microphone array sees three sources in three different directions. Therefore, we have six lines and nine possible intersections of these six lines (Fig. 8 only shows six of the nine possible intersections). The advantage of the ray space image is that it will only ever show three lines resolving the ambiguities raised by the multiple intersections.

4. SPARSE RECOVERY AND RAY SPACE ANALYSIS

There are many possibilities to integrate the sparse recovery and ray space analysis techniques. In [23,24] we apply sparse recovery to the sub-array beamformers in order to enhance the beamforming image for each sub-array and sharpen the ray space image. In this case, the microphone observation matrix, $X(f)$, is limited to the sub-array and the transformation matrix, $A(f)$, corresponds to
the ray space analysis grid. The sparse recovery solution, $\hat{S}_q(n, t, f)$, is then averaged across time frames and summed along each ray to give the imaging result for array $i$:

$$R(i)(m(i), q(i), f) = \sum_{n\in\text{ray}(m(i))} \langle \hat{S}_q(n, t, f) \rangle_i,$$  (19)

where $\langle \cdot \rangle_i$ denotes an average or statistical expectation operation. For further details, please refer to [23].

As an example of the impact that sparse recovery may have on the ray space image, we consider the work detailed in [23]. In that work we used the freely available software package MCRoomSim [25] to simulate a facility for human-robot interaction that is currently under construction. The room simulations include both specular and diffuse reflections. The facility is modelled as a “shoebox” space that spans 6.8 m by 15.5 m by 3.7 m. The surface area and absorption coefficients for the materials lining the walls and ceiling were taken into account when performing the room acoustic simulations ($RT_{60} \approx 0.56$ s). For the numerical simulations, a linear microphone array of 61 microphones with an aliasing frequency of 8,000 Hz was aligned near and parallel to the South wall at a height of 1 m, while 8 source positions were chosen around a 1 m radius circle situated in the centre of the room at a height of 1 m (refer to Fig. 9). Room impulse responses from each source to each of the microphones were calculated.

In a test condition, four simultaneous speech sources were simulated for positions evenly distributed around the circle (positions: 1, 3, 5, 7 in Fig. 9). The speech stimuli consisted of anechoic speech recorded for the Archimedes project [26]. The female talker directivity function provided by MCRoomSim was used. The talkers were oriented facing each other as if around a table. Microphone noise was simulated by adding white Gaussian noise to the microphone signals at a level of $-40$ dB relative to the signal level.

With four simultaneous talkers, a clear and visible improvement is seen for the sparse recovery version of the ray space transform (Fig. 10(b)) compared to the normal ray space transform (Fig. 10(a)). It is possible with the sparse recovery ray space transform to resolve the position of two of the sources.

5. NEURAL NETWORKS

Recently, researchers have investigated using neural networks for direction-of-arrival (DOA) microphone array signal analysis [27–33]. This is somewhat surprising because the DOA problem has generally been considered an analytical problem with well-defined quantities such as the array manifold vector. Nonetheless, neural networks are making some headway and it is interesting to probe what directions this research is moving in and also consider the possible advantages that a neural network may have. The most interesting situations are those that present difficulties to traditional methods. Examples of such situations are when the number of microphones is small or limited [27–29], when working with panoramic arrays in challenging sound conditions [29,30], when there are significant variations and imperfections in the array construction [31], and when working with arrays situated in a multi-room scenario [32].
Early work in this recent research area employs phase information as input to the neural network and generally considers the DOA of only a single source. The phase information could be either the raw phase data [27] or the generalized cross-correlation with phase transform (GCC- PHAT) for various pairs of microphones [32,33]. In [27], a four microphone linear array was used for DOA analysis in reverberant environments. A convolutional neural network (CNN) was trained using the phase component of the short-time Fourier Transform coefficients of the microphone signals as input. The training signals were Gaussian white noise. The performance of the CNN system was robust to small perturbations in the microphone positions and worked under different acoustic conditions. In [28], a CNN was trained to produce high-resolution ray-space images from low-resolution ray-space images. The low resolution images were obtained using a limited number of microphones. Again, the CNN was able to improve the quality of the ray space images. More recent DOA research considers multiple sources and utilizes the full spectral information from the array [29,30]. Interestingly, these researchers explore the full sphere of space using array signals in the spherical harmonic domain. These neural networks are constructed in the form of a convolutional recurrent neural network (CRNN). The researchers indicate that the CRNN architecture performs better than the CNN architecture despite the fact that the source signals are stationary. A convincing explanation for this finding is not provided; perhaps it is because the array is able to pool information over longer time periods. Using four microphones, these approaches achieve angular errors around 5 degrees for a single source and 10 degrees for two sources.

Neural networks may prove to be most useful when it is unclear how to set an analytical optimization criterion for the task at hand. For this consideration, we change directions slightly from the focus of this paper, DOA analyses, and consider beamforming. While there are many analytical objective criteria for beamformers, it is not clear which may perform best given a specific situation. In this light, the research by Xiao et al. [34] is intriguing. In this work, they jointly train two neural network modules: one for beamforming and the other for automatic speech recognition (ASR). Unfortunately, they could not use the latest network architectures for ASR because these did not seem amenable to joint training with full backpropagation computation flowing from the ASR network back to the beamforming network. Nevertheless, the results they obtain using joint training with a cross-entropy acoustic model demonstrate the potential benefits that may be obtained by pursuing this line of research. In summary, the application of neural networks to microphone array signal processing is still in the early phases; there remains much to learn, but it seems a promising approach.

REFERENCES


Craig Jin received a B.S. in Physics from Stanford in 1990, an M.S. in Applied Physics from Caltech, and a Ph.D. from the University of Sydney in 2001. He is currently an Associate Professor within the School of Electrical and Information Engineering at the University of Sydney and is the director of the Computing and Audio Research Laboratory.

Shiduo Yu received a B.E. in Electrical Engineering and M.E in Acoustic Engineering from the Northwestern Polytechnical University, China and is currently working towards a PhD at the University of Sydney.

Fabio Antonacci received a Laurea degree in 2004 in Telecommunication Engineering and a Ph.D. in Information Engineering in 2008, both at Politecnico di Milano, Italy. He is currently an Assistant Professor at Politecnico di Milano.

Augusto Sarti received the M.S. and the Ph.D. degrees in electronic engineering, both from the University of Padua, Italy, in 1988 and 1993, respectively. His graduate studies included a joint graduate program with the University of California at Berkeley. He is currently a Professor at the Politecnico di Milano, Italy.