Sigma-Delta Learning for Super-resolution Source Separation on High-density Microphone Arrays

Amin Fazel and Shantanu Chakrabartty
Department of Electrical and Computer Engineering
Michigan State University
East Lansing, Michigan 48824-1226
Email: {fazel, shantanu}@egr.msu.edu

Abstract—The performance of acoustic source separation algorithms significantly degrades when they applied to signals recorded using miniature microphone arrays where the distances between the microphone elements are much smaller than the wavelength of acoustic signals. This can be attributed to limited dynamic range (determined by analog-to-digital conversion) of the sensor which is insufficient to overcome the artifacts due to large cross-channel redundancy, non-homogeneous mixing and high-dimensionality of the signal space. This paper presents some of the recent progress in the area of sigma-delta learning which integrates statistical learning with analog-to-digital process and enables super-resolution auditory localization and separation. Experiments with synthetic and real recordings demonstrate significant and consistent performance improvements when the proposed approach is used as the analog-to-digital front-end to conventional source separation algorithms.

I. INTRODUCTION

Separation of acoustic sources using miniature microphone arrays [1], [2], poses a significant challenge due to fundamental limitations imposed by the physics of sound propagation [3]. The smaller the distance between the recording elements, the more difficult it is to measure separation cues [3]. An example prototype of a miniature microphone array is shown in Fig. 1 where the distance between the condenser microphone elements is 1cm (approximately 1/34 times the wavelength of sound at a frequency of 1KHz). A traditional approach for separating sources would implement digital signal processing algorithms (for e.g. independent component analysis) subsequent to a quantization (analog-to-digital conversion) process. However, for a miniature microphone array such as the one shown in Fig. 1, a naive quantization of acute auditory and source separation cues could lead to a significant loss of information and hence degradation in performance. In [4] we had proposed a ΣΔ learning technique to overcome this degradation where statistical learning algorithms is directly integrated with the signal measurement process (analog-to-digital conversion). In this paper, we briefly summarize the ΣΔ learning method by highlighting some of its underlying algorithms and by presenting some of the key results which demonstrate the performance improvements that can be expected.

To estimate the performance improvement that can be expected by using the ΣΔ learning approach, consider the framework of a conventional source separation algorithm which is shown in Fig. 2(a). The “analog” signal \( \mathbf{x} \in \mathbb{R}^M \) recorded at each of the sensor array is given by \( \mathbf{x} = \mathbf{As} \), \( \mathbf{A} \in \mathbb{R}^{M \times M} \) being the mixing matrix and \( \mathbf{s} \in \mathbb{R}^M \) being the independent sources of interest. This simplified linear model has been shown to be applicable under far-field recording conditions which is almost always the case for miniature microphone arrays [6], [2]. In a conventional source separation framework shown in Fig. 2(a), the recorded signals are first digitized and then processed by a digital signal processor (DSP) which implements an un-mixing operation. The ideal unmixing matrix is given by \( \mathbf{W} = \mathbf{A}^{-1} \) which is then used to recover the source signals \( \mathbf{s} \in \mathbb{R}^M \). The effect of quantization in this conventional approach is

\[
\tilde{\mathbf{s}}_d = \mathbf{W}(\mathbf{x} + \mathbf{q}) = \mathbf{s} + \mathbf{A}^{-1}\mathbf{q}
\]

where \( \mathbf{q} \) denotes the additive quantization error introduced during the digitization process. The reconstruction error between the recovered signal \( \tilde{\mathbf{s}}_d \) and the source signal \( \mathbf{s} \) can be expressed as

\[
\|\tilde{\mathbf{s}}_d - \mathbf{s}\| \leq \|\mathbf{A}^{-1}\| \|\mathbf{q}\|
\]

which indicates that under ideal reconstruction conditions, the performance of conventional source separation algorithm is...
signal now can be expressed as \( \tilde{D} P \). For source separation \( \tilde{D} \) case, the reconstructed signal \( \tilde{P} \) by signals recorded by the microphone array is first transformed in the analog domain before being quantized. In this section, we describe only the key steps in the super-resolution source separation algorithm; section 3 presents results from numerical simulations and in the following sections we show how \( \Sigma \Delta \) learning can adaptively determine the transform \( P \) and deliver performance improvement as predicted by equation (5). The paper is organized as follows: Section 2 describes the super-resolution source separation algorithm; section 3 presents results from far-field microphone arrays. Section 5 concludes the paper with some final remarks.

\[ ||s_m - s|| = ||A^{-1}P^{-1}q|| \leq ||A^{-1}P^{-1}|| \cdot ||q||. \]  

Therefore, the reconstruction error can now be determined by the analog projection matrix \( P \), mixing matrix \( A \), and quantization additive error \( q \). Adapting matrix \( P \) in a way that ensures input signal normalization in analog domain i.e.,

\[ ||A^{-1}P^{-1}|| = 1 \]  

reduces the equation (3) to \( ||s_m - s|| \leq ||q|| \). Employing the framework in Fig. 2(a) over the conventional source separation framework, the expected performance improvement can be expressed as

\[ PI = -20 \log ||A^{-1}|| \]  

which could be significant when the mixing is near singular, as is the case for miniature microphone arrays. Visually, this performance improvement can be appreciated by an image processing example shown in Fig. 3. The reference image (with the size of 128 x 128 pixels) shown in Fig. 3(a) is transformed by a near-singular linear transformation \( A \) (with a condition number of \( 2^{13} \)) whose rasterized form is shown in Fig. 3(b). The result using the traditional quantization followed by reconstruction (using \( A^{-1} \)) approach is shown in Fig. 3(c) where as the result using quantization followed by reconstruction is shown in Fig. 3(d). The quantization and reconstruction artifacts can be clearly seen in the traditional case which is not evident in the "smart" reconstruction procedure as prescribed by equation (4). In the above example we have assumed prior knowledge of \( A^{-1} \) and in the following sections we show how \( \Sigma \Delta \) learning can adaptively determine the transform \( P \) and deliver performance improvement as predicted by equation (5). The paper is organized as follows: Section 2 describes the super-resolution source separation algorithm; section 3 presents results from numerical simulations and section 4 presents results from far-field microphone arrays. Section 5 concludes the paper with some final remarks.

II. SUPER-RESOLUTION \( \Sigma \Delta \) LEARNING ALGORITHM

In this section, we describe only the key steps in the \( \Sigma \Delta \) learning algorithm. For details about the derivation and the properties of the algorithm, the readers are referred to [4], [5]. Given an \( M \) dimensional input vector \( x \in \mathbb{R}^M \) and an internal state vector \( v \in \mathbb{R}^M \), the \( \Sigma \Delta \) learning algorithm estimates parameters of the linear transformation \( P = AB \) in equation (4) according to the following recursions where \( n \) denotes a discrete time index:

\[ d[n] = Q[v[n - 1]] \]  

\[ v[n] = v[n - 1] + B[n]{x[n]} - \Lambda^{-1}d[n] \]  

\[ B[n] = B[n - 1] - 2^{-P}\phi(d[n])d[n]T; \quad B[n] \in C \]  

\[ \lambda_i = \max(||B[n]{x[n]}||); N_1 > n > N_0 \]  

\[ B[n] \in \mathbb{R}^M \times \mathbb{R}^M \] is a linear transform satisfying constraint \( C \) and \( \Lambda \in \mathbb{R}^M \times \mathbb{R}^M \) is a diagonal matrix. \( B \) in equation (9) is a parameter which determines the resolution of updates.
the parameter matrix $B$. $Q$ in equation (7) is a quantization function that converts the analog state vector $v[n-1]$ into a digitized representation. An example of a quantization function could be $Q(.) = \text{sgn}(.)$ which is a two step quantizer. The function $\phi(.)$ in the update (9) could be any monotonic, non-linear, bounded function, for e.g., $\phi(.) = \tanh(.)$. The main concept of this approach is depicted in Fig. 2(b) with block diagrams in Fig. 2(c).

Under the assumption that the recursions are stable (the readers are referred to [4], [5] for details) implying $B[n] \xrightarrow{n \to \infty} B^*$, then the recursions (7)- (9) lead to the following results:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} d[n] = \lim_{N \to \infty} \Lambda B^* x[n] \quad (10)$$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \phi(d[n])d[n]^T = 0 \quad (11)$$

Equation (10) indicates that the time-domain average of the quantized vector signal $d[n]$ converges to the time-domain average of the transformed input vector $x[n]$. The second equation (11) indicates that the quantized outputs $d[n]$ are de-correlated with respect to each other. Because the choice of $\phi(.)$ and the quantizer $Q(.)$ is arbitrary, equations (7)- (9) also leads to outputs which are non-linearly de-correlated. The de-correlation matrix is then given by $P = \Lambda B^*$.

### III. Numerical Evaluation

In this section we verify the performance improvement of the proposed $\Sigma \Delta$ learning over the standard $\Sigma \Delta$ modulator as predicted by the equations (5). For this controlled experiment we have used the following two synthetic time series:

$$s_1(t) = 500t - [500t]$$
$$s_2(t) = \sin(1000t + 10\cos(100t)) \quad (12)$$

These two signals were mixed using a random ill-conditioned matrix $A$. The two-dimensional mixed signals then processed by two modulator: standard $\Sigma \Delta$ and the proposed $\Sigma \Delta$ learner. The outputs of each of these modulators were then used to reconstruct the source signals according to

$$\tilde{s}_d = A^{-1}x \quad (13)$$
$$\tilde{s}_m = A^{-1}B^{-1}A^{-1}x \quad (14)$$

assuming that the un-mixing matrix $A^{-1}$ can be perfectly determined. The equations (13) and (14) represent the following two cases: (a) $\tilde{s}_d$ which is the signal reconstructed using a $\Sigma \Delta$ modulator without any learning (denoted by without); (b) $\tilde{s}_m$ which is the signal reconstructed using $\Sigma \Delta$ learning with resolution enhancement (denoted by with). To quantify this improvement, we compared the signal-to-error ratio (SER) for these two cases which is defined as

$$SER = \log_2 \left\{ \frac{||s||_2}{||s - \tilde{s}||_2} \right\} \quad (15)$$

where $s$ and $\tilde{s}$ is based on the definition in (13) and (14). To compute the mean SER and its variance 10 different ill-conditioned mixing matrices $A$ with a fixed condition number of $2^{12}$ were chosen and the mean/variances were calculated across different experimental runs. Fig. 4(a) shows the performance comparison for different values of over-sampling ratio (OSR). It can be seen in Fig. 4(a) that as the OSR of the $\Sigma \Delta$ modulator increases, the SER increases. This is consistent with results reported for $\Sigma \Delta$ modulators where OSR is directly related to the resolution of the “analog-to-digital” conversion. However, it can be seen that for all conditions of OSR, $\Sigma \Delta$ learning with resolution enhancement outperforms the standard $\Sigma \Delta$ modulator. Figure 4(b) compares the performance of $\Sigma \Delta$ learner with a standard one when the condition number of the ill-conditioned mixing matrix is varied for fixed OSR of 256. The results again show that $\Sigma \Delta$ learner with resolution enhancement demonstrates consistent performance improvement over the traditional $\Sigma \Delta$ modulator. Also, as expected the SER performance for both cases deteriorates with the increasing condition number, which indicates that the mixing becomes more singular.

### IV. Far-field Experiments and Results

In this setup, we have used the $\Sigma \Delta$ learning modulator in a real recording speech signals from a prototype miniature microphone array, similar to the set up described in [2]. The four omni-directional microphones were place in a circular array with radius 0.5 (similar to prototype shown in Figure 1). The speech signals were presented through loudspeakers from a distance of 1.5 m to the array where the data was recorded and archived. Figure 5(a) and (b) shows the spectrogram of the speech signals recorded from the microphone array. The two spectrograms look similar, thus emulating a “near-far” recording scenario where a dominant source masks the background weak source. The archived speech signals then presented to two source separation frameworks using SOBI as the DSP based algorithm where one of them used the
standard ΣΔ modulator and the other one used ΣΔ learning. Figure 5(c) and (d) show the spectrogram of the separated speech signals obtained from the framework used standard ΣΔ modulator. Figure 5(e) and (f) show the spectrogram of the separated speech signals obtained with ΣΔ learning. A visual comparison of the spectrograms show that separated speech signal without ΣΔ learning contain more quantization artifacts which can be seen as the broadband noise in Fig. 5(c) and (d).

V. CONCLUSION AND DISCUSSION

In this paper, we have argued that the conventional source separation techniques which consist of quantization followed by DSP based algorithms fail to deliver robust performance when they are applied to the signals recorded by high-density microphone arrays. We reviewed our previously reported results in ΣΔ learning which integrates statistical learning with analog-to-digital conversion and hence can be used for designing “smart” ADCs. One of applications of this integration is the ability to exploit spatial correlations to resolve acute differences in signals recorded using miniature/compact microphone array. Even though in this paper we have used ΣΔ learning for speech based source separation algorithms, there are other high-density arrays that can benefit from this technique including microelectrode arrays and micro-antenna arrays, as shown by the visual example in Fig. 3. The experimental results presented in this paper demonstrated that the ΣΔ learning algorithm consistently led to a superior performance over the classical approach when applied for source separation in high-density microphone arrays.

REFERENCES