The Journal of the Acoustical Society of America

A blind source separation approach for humpback whale song separation

--Manuscript Draft--

Manuscript Number:	JASA-01084
Full Title:	A blind source separation approach for humpback whale song separation
Short Title:	Separation of humpback whale song
Article Type:	Regular Article
Section/Category:	Signal Processing in Acoustics
Keywords:	bioacoustic signal processing; blind source separation; humpback whale song; noise reduction.
Abstract:	Many marine mammal species are highly social and are frequently encountered in groups or aggregations. When conducting passive acoustic monitoring in such circumstances, recordings commonly contain vocalizations of multiple individuals which overlap in time and frequency. This paper considers the use of blind source separation as a method for processing these recordings to separate the calls of individuals. The example problem considered here is that of the songs of humpback whales. The high levels of noise and long impulse responses can make source separation in underwater contexts a challenging proposition. The approach is based on time-frequency masking, allied to a noise reduction process. The technique is assessed using simulated and measured data sets.

A blind source separation approach for humpback whale song separation

Zhenbin Zhang* and Paul R. White

Institute of Sound and Vibration Research,

University of Southampton, Southampton, SO17 1BJ, UK

(Dated: September 15, 2016)

Abstract

Many marine mammal species are highly social and are frequently encountered in groups or aggregations. When conducting passive acoustic monitoring in such circumstances, recordings commonly contain vocalizations of multiple individuals which overlap in time and frequency. This paper considers the use of blind source separation as a method for processing these recordings to separate the calls of individuals. The example problem considered here is that of the songs of humpback whales. The high levels of noise and long impulse responses can make source separation in underwater contexts a challenging proposition. The approach is based on time-frequency masking, allied to a noise reduction process. The technique is assessed using simulated and measured data sets.

PACS numbers: PACS: 43.60.Fg, 43.30.Sf, 43.60.Hj

* zhangzhenbin737@hotmail.com; Corresponding author.

I. INTRODUCTION

Acoustics is one of the most effective methods for monitoring underwater environments. In particular, marine mammals rely primarily on acoustics for interacting with the environment and their conspecifics. This is because sound represents the most effective method to transmit information underwater [1]. The highly social nature of many marine mammal species mean that they are frequently encountered in groups [2]. Consequently acoustic recordings of marine mammals frequently contain vocalizations from more than one individual, these vocalizations typically occur simultaneously and in the same frequency band. This overlap means that trivial processing operations like filtering or time-gating will not, in general, separate these calls. The approach considered in this work is based on the concept of Blind Source Separation [3] which provides a method for separating vocalizations based on recordings from more than one sensor, without knowledge of the sensor geometry or models of propagation conditions. However, the long reverberation times, relatively high noise level, and large number of sources that may be observed at a particular time, means that separating individual source remains a challenge. In this paper, we aim to separate humpback whale songs recorded in the St Marie channel, Madagascar. This is a breeding ground in which a high density of male singers is encountered [4]. The objective is to take a recording containing the songs from multiple singers, and decompose it into songs from individual singers.

The humpback whale (*Megaptera novaeangliae*) is a species of baleen whale. They undertake annual migration from warm waters where they reproduce to their feeding grounds in colder waters [5]. The complex and mysterious songs that they produce have attracted marine biologists for decades [6–8]. Humpback whale songs are long cyclical sequences produced by males especially in the breeding season [9]; while the social sounds are produced by both males and females in social interactions such as feeding [2]. The purpose of their complicated songs is still not clear, though it is assumed to relate to reproduction [10].

The structure of humpback whale songs has been defined as a hierarchical series of units, phrases and themes [9]. The building block of a song is a unit which is defined as a continuous sound between two silences. The duration of units varies considerably, with some units lasting less than 1 s and other units extending to 5 s. Units have one of several acoustic structures; some are based on frequency modulated narrow band components, whilst~~,~~ other units are a series of pulsed sounds [7]. In some cases, one unit can consist of two or more sounds, each discrete sound being called a subunit [13]. Two or more units can be repeated in a specific pattern to form a phrase, and phrases are combined and repeated several times to form a theme, finally a song is composed of several distinct themes [6, 9]. A series of songs within which there is no pause longer than one minute is termed a song session [9].

The classification of songs into different units has been conducted by scientists in the attempt to understand the function of humpback whale songs and their evolution [11–13]. Humpback whale songs are recorded through hydrophones and the quality of the recording is crucial for the subsequent research [11]. There are at least two commonly employed ways to obtain recordings of humpback songs. One way is taking hydrophone in a boat and when the whale position is identified, the hydrophone is lowered into the water to record the song [13]. The other way is using a moored hydrophone array [14]. Moored arrays have the advantage of being less affected by inclement weather and have the ability ~~of~~ to monitor for extended periods, but their cost and problems associated with siting and maintaining them hampers their use. In many analyses, such as for automatic classification of songs [11–13], a recording containing a single singer is required. In some locations, it is challenging to find an isolated singer and make a recording without interference from vocalisations from other whales. In such conditions, recordings commonly contain a mixture of multiple songs, which can cause great difficulty for subsequent analyses [11]. Therefore, the ability to separate humpback songs potentially provides a powerful tool for the analysis of the song characteristics.

In this study, the data analysed is collected from a fixed hydrophone array located off the island of St Marie, in Madagascar. The high density of singers in this location means that, for the vast majority of the time, multiple singers can be heard simultaneously. Our goal is to automatically separate the mixtures and obtain a recording of individual songs. In previous research, visual inspection of song spectrograms and empirically listening to the vocalizations were utilized in order to find sections of recordings containing only one singer [2, 12, 14]. These manual analyses can be extremely time consuming and raise serious concerns regarding subjectivity [2].

Source separation methods are techniques in which one can take a set of recordings of mixtures of multiple sources and from them construct estimates of the original source signal [3]. Blind source separation (BSS) methods are a subset of source separation methods in which the propagation paths are unknown and they rely only broad assumptions about the source characteristics to achieve separation. BSS problems which are characterised by a non-trivial impulse response functions are referred to as convolutive. In particular, in shallow water environments the long reverberation times, which characterise the complex underwater acoustic, impede the performance of BSS methods. The problem is made even more challenging in many circumstances since it is under-determined, i.e. the number of sources exceeds the number of channels of data. Previously, the BSS approach has been utilized to enhance marine mammal vocalizations, where a two channel second order statistics (SOS) based BSS ~~is~~ developed for enhancing manatee vocalizations [15].

Various BSS methods which solve convolutive and under-determined problems have been proposed. The performance of these methods has been compared as part of the signal separation evaluation campaign SiSEC [16]. Such methods include~~,~~ model-based EM source separation and localization (MESSL), which uses a probability model of interaural parameters to ~~allocate~~ each spectrogram point into a cluster, with each cluster corresponding to a different source [17]. The restriction of this algorithm is that it is built on a model of the human auditory binaural responses and relies on prior information of human auditory system. An alternative approach is based on a set of full-rank spatial covariance model [18]. In this method the contribution of each source to all mixture channels in the time-frequency domain is modelled as a zero-mean Gaussian random variable, whose covariance encodes the spatial characteristics of the source. The drawback of this method is the sensitivity of estimation of spatial covariance and source variance. A further, popular approach, is based on Nonnegative matrix factorization (NMF), which models each source with a complex valued tensor and eight nonnegative matrices [19]. It has been demonstrated that the NMF model is more suitable to music than speech [19], whilst the vocalisations we are interested in are referred to as song, their structure rather more closely corresponds to that of speech than music.

The BSS algorithm selected for this study is based on separation in the time-frequency domain [20], which we refer to as the Sawada algorithm. In this method, the separation process is based on the short-time Fourier transform (STFT) and is divided into two stages. In the first stage, referred to as bin-wise clustering. For every frequency bin the samples representing the measurements of the mixture~~,~~ are clustered, using a form of Expectation Maximization (EM) algorithm [21]. Each cluster represents the samples from one of the sources. The second stage is to solve the~~,~~ so-called~~,~~ permutation problem, to ensure that source components in different frequency bins are associated with each other correctly [22].

This paper is organized as follows. Section II provides an overview of the Sawada algorithm for convolutive under-dermined BSS. The effectiveness and robustness of the Sawada method for humpback songs separation are verified through separation of artificial mixtures of humpback song generated by the underwater impulse response model discussed in section III. The separation of real world signals is addressed in section IV, where a novel approach to noise reduction based on a weighted median threshold scheme applied to the Sawada method. Finally, conclusions are presented in section V.

II. THE SAWADA ALGORITHM

A. Signal Notation

Let $\mathrm{s}_1(t), \cdots, \mathrm{s}_N(t)$ represent the set of N source signals and $\mathrm{x}_1(t), \cdots, \mathrm{x}_M(t)$ be the M hydrophone observations. The observations $\mathrm{x}_j(t)$ at the j^{th} hydrophone can be described as a sum of source images,

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \mathrm{s}_{jk}^{\mathrm{img}}(t), \tag{1}$$

where the source images $\mathrm{s}_{jk}^{\mathrm{img}}(t)$ represent the signal received from the k^{th} source on the j^{th} hydrophone. The image sources are represented as the convolution of the k^{th} source with the corresponding impulse response, $\mathrm{h}_{jk}(t)$, from the source to the j^{th} hydrophone. Thus the source images can be expressed as,

$$\mathrm{s}_{jk}^{\mathrm{img}}(t) = \sum_{l} \mathrm{h}_{jk}(l)\mathrm{s}_k(t - l). \tag{2}$$

The goal of blind separation is to obtain sets of separated signals $\{\mathrm{y}_{11}, \cdots, \mathrm{y}_{1M}\}, \cdots, \{\mathrm{y}_{N1}, \cdots, \mathrm{y}_{NM}\}$, such that $\mathrm{y}_{jk}(t)$ is an estimate of the source image $\mathrm{s}_{jk}^{\mathrm{img}}(t)$, only using the information from the set of observed mixtures $\mathrm{x}_1(t), \cdots, \mathrm{x}_M(t)$.

B. Transform into frequency domain

Assuming the observations (1) are sampled at a frequency f_s, the algorithm begins by converting these signals into frequency domain time series signals $x_j(\tau, f)$ by an STFT, defined as follows,

$$x_j(\tau, f) = \sum_{t'=0,t_s,\cdots,(L-1)t_s} w(t')\mathrm{x}_j(t'+\tau)e^{-i2\pi f t'}, \tag{3}$$

where $t_s = 1/f_s$ is the sampling interval, L is the number of samples in the spectral window, $w(t)$. S is the window shift in samples, $\tau = 0, St_s, 2St_s, \cdots$ is the starting time of each frame, T is the total number of samples, $f = 0, (1/L)f_s, \ldots, ((L-1)/L)f_s$ is the frequency index. $w(t)$ can be any suitable windowing function (here the Hanning window is used).

Suppose that the frame size L is long enough to cover the majority of the impulse responses h_{jk}, then the convolutive mixture model (2) can be approximated as an instantaneous mixture model and expressed in matrix form for each frequency as [17, 18, 20]

$$\mathbf{x}(\tau, f) = \sum_{k=1}^{N} \mathbf{h}_k(f)s_k(\tau, f) + \mathbf{n}(\tau, f), \tag{4}$$

where $\mathbf{h}_k(f)$ is an $M \times 1$ column vector of the Fourier transform of $\mathrm{h}_{jk}(t)$, $j = 1, \cdots, M$ and $\mathbf{n}(\tau, f)$ is an $M \times 1$ column vector of noise components that includes both additive background noise and unresolved reverberant components.

The Sawada algorithm, like several other BSS methods, relies on the W-disjoint property of the signals [23]. This requires that signals are sparse such that their STFTs do not have significant overlap, specifically every time-frequency cell is dominated by a single source. This allows one to simplify the mixture model (4), so that, for some k, one can express (4) as [20]:

$$\mathbf{x}(\tau, f) = \mathbf{h}_k(f)s_k(\tau, f) + \tilde{\mathbf{n}}(\tau, f). \tag{5}$$

Note that the notation for the noise term has be modified to reflect that it now also contains residual components of other sources in time-frequency slots. These additional residual components are a consequence of the fact that, in practice, the W-disjoint property only holds in an approximate fashion. To construct an estimate of the k^{th} signal on the j^{th} hydrophone, a binary mask, $M_k(\tau, f)$ is applied to the STFT of $x_j(t)$ as follows

$$y_{kj}(\tau, f) = M_k(\tau, f)x_j(\tau, f). \tag{6}$$

At the end of the processing, the time-domain separated signals $\mathrm{y}_{kj}(t), k = 1, \cdots, N, j = 1, \cdots, M$ are calculated using an inverse STFT applied to the separated components defined in (6).

C. Frequency bin-wise clustering

Separation is achieved through clustering of the data in each frequency bin in the STFT [20]. Separate clusters are formed for each source. The clustering can be performed according to the information in the vector $\mathbf{x}(\tau, f)$. Based on (5), assuming the noise term $\tilde{\mathbf{n}}(\tau, f)$ is negligible. It is possible to eliminate the effect of source amplitude $s_k(\tau, f)$ from $\mathbf{x}(\tau, f)$, by normalizing it to have a unit norm [20]

$$\check{\mathbf{x}}(\tau, f) = \frac{\mathbf{x}(\tau, f)}{\|\mathbf{x}(\tau, f)\|} = \frac{s_k(\tau, f)}{|s_k(\tau, f)|} \frac{\mathbf{h}_k(f)}{\|\mathbf{h}_k(f)\|}. \tag{7}$$

The next step is to perform pre-whitening on the normalized observation vectors $\check{\mathbf{x}}$, since that makes the clustering more robust. The whitening can be expressed as

$$\bar{\mathbf{x}}(\tau, f) = \mathbf{V}\check{\mathbf{x}}(\tau, f). \tag{8}$$

where $\bar{\mathbf{x}}(\tau, f)$ are the whitened observations. The whitening matrix $\mathbf{V}$ is calculated as $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^H$ where $\mathbf{D}$ and $\mathbf{E}$ are obtained from the eigenvalue decomposition $\mathrm{E}\left\{\check{\mathbf{x}}\check{\mathbf{x}}^H\right\} = \mathbf{E}\mathbf{D}\mathbf{E}^H$ of the correlation matrix. Super-script H denotes Hermitian transpose. The normalisation procedure (7) is reapplied after the pre-whitening process, to produce the conditioned data vector $\widehat{\mathbf{x}}$.

The clustering is performed on each frequency bin individually, assuming that the number of sources, i.e. the number of clusters, N is known. The vectors $\widehat{\mathbf{x}}(\tau, f)$ for all time-frequency cells (τ, f) are clustered into N classes $C_1, \cdots, C_N$, each ~~of~~ cluster corresponds to a source signal s_k. The posterior probability $P(C_k|\mathbf{x})$, a measure of how likely the vector $\mathbf{x}(\tau, f)$ is to belong to the k^{th} class, is calculated. The clustering algorithm is based on the line orientation idea utilized within the LOST algorithm [24] and employs a complex probability density function of the form

$$p(\mathbf{x}|\mathbf{a}_i,\sigma_i)=\frac{1}{(\pi\sigma_i^2)^{M-1}}\exp\left(-\frac{\left\|\widehat{\mathbf{x}}-(\mathbf{a}_i^H\widehat{\mathbf{x}})\mathbf{a}_i\right\|^2}{\sigma_i^2}\right), \tag{9}$$

where $\mathbf{a}_i$ is the unit-norm centroid and σ_i^2 is the variance, which depicts the extent by which the data deviates from the centroid. $(\mathbf{a}_i^H\widehat{\mathbf{x}})\mathbf{a}_i$ represents the orthogonal projection of $\widehat{\mathbf{x}}$ onto $\mathbf{a}_i$. The density function models the distances between the data points and the subspace spanned by $\mathbf{a}_i$. The term $\left\|\widehat{\mathbf{x}}-(\mathbf{a}_i^H\widehat{\mathbf{x}})\mathbf{a}_i\right\|^2$ is zero if $\widehat{\mathbf{x}}=\mathbf{a}_i$. In this case, the observation vector and centroid overlap with each other, so $p(\mathbf{x}|\mathbf{a}_i,\sigma_i)$ achieves its maximum value. The joint density function, $p(\mathbf{x})$ is given by the mixture model,

$$p(\mathbf{x}|\theta)=\sum_{i=1}^{N}\alpha_i p(\mathbf{x}|\mathbf{a}_i,\sigma_i), \tag{10}$$

with parameters $\theta=\mathbf{a}_1,\sigma_1,\alpha_1,\cdots,\mathbf{a}_N,\sigma_N,\alpha_N$. The mixture ratios α_i satisfy $\alpha_1+\cdots+\alpha_N=1$ and $0\leqslant\alpha_i\leqslant 1$. The prior distribution for these mixture ratios is modelled by the Dirichlet distribution

$$p(\alpha_1,\cdots,\alpha_N)=\frac{\Gamma(N\phi)}{\Gamma(\phi)^N}\prod_{i=1}^{N}\alpha_i^{(\phi-1)}, \tag{11}$$

where ϕ is a hyper-parameter controlling the width of the Dirichlet prior.

The parameters θ are estimated using the EM algorithm [21]. The EM algorithm provides a general iterative approach for computing maximum likelihood estimates. The main advantage of the EM algorithm is that it often allows treatment of difficult maximum likelihood problems containing a large number of parameters and characterised by a highly non-linear likelihood functions in terms of a sequence of simpler maximization problems [3].

In the E-step, the posterior probabilities are calculated using

$$P(C_i|\mathbf{x},\theta')=\frac{\alpha' p(\mathbf{x}|\mathbf{a}_i',\sigma_i')}{p(\mathbf{x}|\theta')}=\frac{\alpha' p(\mathbf{x}|\mathbf{a}_i',\sigma_i')}{\sum_{i=1}^{N}\alpha' p(\mathbf{x}|\mathbf{a}_i',\sigma_i')} \tag{12}$$

with the current parameter set

$$\theta'=\left\{\mathbf{a}_1',\sigma_1',\alpha_1',\cdots,\mathbf{a}_N',\sigma_N',\alpha_N'\right\}. \tag{13}$$

In the M-step, the log-likelihood function is maximised by maximising $Q(\theta, \theta') + \log p(\theta)$, where,

$$Q(\theta, \theta') = \sum_{\tau}^{T} \sum_{i=1}^{N} P(C_i | \mathbf{x}(\tau), \theta') \log \alpha_i p(\mathbf{x}(\tau) | \mathbf{a}_i, \sigma_i). \tag{14}$$

The posterior probabilities, (12), are used to compute the masks in (6) via:

$$\boldsymbol{M}_k = \begin{cases} 1, & \text{if } P(C_k|\mathbf{x}) \geq P(C_{k'}|\mathbf{x}), \forall k' \neq k \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

D. Permutation Alignment

Once clustering has been completed in each frequency bin then one has an estimate of which data points belong to each of the sources. However, the ordering of sources in each frequency bin is arbitrary. So that in two frequency bins there is no reason to assume that, say, the first source in one bin corresponds to the same physical source as the first source in another frequency bin. In the permutation alignment stage, the algorithm aims to ensure that source identities are preserved between frequency bins. This method uses the sequence of posterior probabilities to resolve this ambiguity and align the sources between adjacent frequency bins. This is based on the fact that the time sequences of the posterior probabilities for one source in different frequency bins tend to be highly correlated, implying that the source tends to be active at the same time across a range of frequency bins [20]. This is consistent with the structure of humpback whale songs. Since a vocalising whale will emit a sequence of unit, and each unit will typically contain multiple frequencies, producing a pattern of activation which is well correlated across frequency.

As a consequence~~s~~ of the clustering step, one has ~~estimation~~ of the posterior probabilities $P(C_i|\mathbf{x}(\tau, f))$ for $i = 1, \cdots, N$, see (12) at all time-frequency slots (τ, f). However, the class order $C_1, \cdots, C_N$ may be different from one frequency to another. So there is a need to reorder the indices so that the same index corresponds to the same source over all frequencies. In other words, we need to identify the correct permutation for all frequencies f,

$$\Pi_f := \{1, \cdots, N\} \rightarrow \{1, \cdots, N\}. \tag{16}$$

Due to the fact that posterior probability sequences that belong to the same source generally have similar temporal patterns at different frequencies. The permutation Π_f can be deduced by examining the correlation coefficients between two posterior probabilities sequences at different frequencies [20],

$$\rho(\nu_i, \nu_j) = \frac{E\left\{(\nu_i - \mu_i)(\nu_j - \mu_j)\right\}}{\sigma_i \sigma_j}, \tag{17}$$

where ν_i and ν_j are two different posterior probabilities sequences corresponding to different frequency bins, $\mu_i = E\left\{\nu_i\right\}$ is the mean and σ_i is the standard deviation of ν_i. The correlation coefficient of any two sequences is bounded by $-1 \leqslant \rho(\upsilon_i, \upsilon_j) \leqslant 1$, and becomes 1 if two sequences are identical up to a positive scaling and an additive offset.

In order to reduce the computational load that aligning a large number of frequency bins can impose, a strategy based on a rough global optimization followed by a fine local optimization is utilized [25]. In the rough optimization stage, the correlation is calculated with respect to the average posterior probabilities. In the fine optimization procedure, under the assumption that the signals should have a harmonic structure, the correlation is calculated between adjacent frequencies and harmonic frequencies.

III. SEPARATION OF ARTIFICIALLY GENERATED HUMPBACK SONG

A. Underwater impulse response generation

To test the effectiveness of this method for the separation of humpback songs, we first conducted tests based on simulated data, employing recorded humpback whale songs and modeled acoustic impulse responses. The humpback songs were recorded by a hydrophone deployed from a small boat in the St Marie channel, Madagascar during 2012. The hydrophone was deployed and when an isolated humpback singer was located, the song was then recorded for as long as the vessel remained in the vicinity of the singer. The recording used here is of high quality with a good Signal to Noise Ratio (SNR) as the hydrophone was close to the singer. The sampling frequency of the original recordings was 48 kHz, but for the simulations the data was down-sampled to 8 kHz since there is little acoustic energy above 4 kHz.

The simulated configuration is illustrated in Fig 1, where x_1, x_2 denote two receivers, s_1, s_2, s_3 denote three sources. This configuration is designed to capture the essence of the

FIG. 1. The simulation experimental configuration for sources and receivers in the underwater environment. x_1, x_2 represent two receivers, and s_1, s_2, s_3 denote three sources.

FIG. 2. The impulse responses for both the source and the receiver located depth 50 m and their distance apart is 510 m in the left panel and 3131 m in the right panel. The origin of the time axis is defined so that the majority of the propagation delay has been removed.

physical environment in which real signals are recorded. The two receivers are configured so they are the same distance, d, from central junction point, o. r_k, denotes the distance between the source, s_k, and the centre of the array, and θ_k denotes the angle relating the source, s_k, to the array as shown in the figure. $d = 150$ m, $r_1 = 300$ m, $r_2 = 500$ m, $r_3 = 400$ m, $\theta_1 = 30°$, $\theta_2 = 120°$, $\theta_3 = 90°$, the depth of both sources and receivers is 50 m, while the water depth is 100 m.

The impulse response from the sources to the receivers is generated using a model of the underwater environment. The Underwater Acoustic Propagation Modelling software-AcTUP V2.2L [26] was adopted to generate the impulse response. The fully range dependent parabolic equation code for fluid seabeds (RAMGeo) model was employed. The marine sediment is assumed to be sand. The compressional velocity and the shear wave velocity of sand are set as 1798 m/s and 160 m/s respectively[27]. The compressional wave absorption is set as 0.77 dB/wavelength, while the shear wave absorption is 3.6 dB/wavelength [28, 29]. For a fixed position of source, the frequency response in different receiver position was obtained through RAMGeo model. The impulse response is calculated via inverse Fourier transform of the frequency response. Two examples of impulse responses are shown in Fig 2, in which the source and the receiver are 510 m apart in the left panel, and 3131 m apart in the right panel. Their reverberation times (T60s) are 0.4 s and 1.5 s respectively.

B. Source separation evaluation

The separation performance is evaluated using the Signal to Distortion Ratio (SDR) [30]. The estimates $\hat{\mathrm{s}}_{jk}^{img}(t)$ of the spatial signals of all sources k are compared with the true source image signals. An estimated source image can be decomposed as

$$\hat{\mathrm{s}}_{jk}^{img}(t) = \mathrm{s}_{jk}^{img}(t) + \mathrm{e}_{jk}^{spat}(t) + \mathrm{e}_{jk}^{interf}(t) + \mathrm{e}_{jk}^{artif}(t), \tag{18}$$

where $\mathrm{s}_{jk}^{img}(t)$ is the true source image and $\mathrm{e}_{jk}^{spat}(t)$, $\mathrm{e}_{jk}^{interf}(t)$, $\mathrm{e}_{jk}^{artif}(t)$ are distinct error components representing spatial (or filtering) distortion, interference and artifacts respectively. They are computed by least-squares projection of the estimated source image onto the corresponding signal subspace. The total error is defined by the power ratio between wanted and unwanted components

$$SDR_k = 10\log_{10}\frac{\sum_{j=i}^{M}\sum_t \mathrm{s}_{jk}^{img}(t)^2}{\sum_{j=i}^{M}\sum_t(\hat{\mathrm{s}}_{jk}^{img}(t) - \mathrm{s}_{jk}^{img}(t))^2}. \tag{19}$$

Greater values for the SDR represent better separation performance.

For the purposes of comparison, we introduce the concept of ideal time-frequency binary masks [23], which are designed using the source information by

$$M^{ideal}{}_k = \begin{cases} 1, & \text{if } \sum_j \left|\mathrm{s}_{jk}^{img}\right|^2 \geqslant \sum_j \left|\mathrm{s}_{jk'}^{img}\right|^2, \forall k' \neq k. \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

The ideal binary masks can only be constructed in a simulation environment where the true source signals and the impulse responses from sources to receivers are known. The separation performance obtained through use of the ideal binary masks provides a useful upper bound on the performance one can expect from a practical BSS method based on binary time-frequency masking.

C. Separation results using the Sawada method

The Sawada method was employed to blindly separate multiple sources in the simulated underwater environment. In this example, three sources and two receivers are utilized, so the problem is under-determined, as shown in figure 1. The geometry selected was to use $r_1 = 300$ m, $r_2 = 500$ m, and $r_3 = 400$ m, with bearing $\theta_1 = 30°$, $\theta_2 = 120°$, $\theta_3 = 90°$. The sources are 30 s humpback songs resampled to a rate of 8 kHz. The FFT size and the window shift sample size in the STFT are chosen to be 4096 and 1024 respectively, which correspond to 2 Hz frequency resolution in the STFT and 125 ms temporal increment. These values were selected as they yielded the best SDRs.

The choice of the FFT size is an important parameter controlling the performance of the separation algorithm. There are several factors that need to be considered when selecting it. Firstly this algorithm, like many of this kind, relies on approximating convolution in the time domain as multiplication in the frequency domain see (4). For this to hold in a digital context with finite duration windows, ~~then~~ the FFT size needs to be large compared to the duration of the impulse response. The second factor that needs to be considered is that long FFTs result in fewer time samples in any one frequency bin (assuming a fixed duration of a recording and fixed hop size). This provides fewer data samples in the clustering stage, so degrading its performance. A third factor that needs to be considered is the fact that the sparseness of the humpback whale songs in the STFT domain reduces if the FFT size becomes too large. Typically, in a song, the units are separated by pauses of roughly 1 s duration, if the FFT size exceeds this then the intervals between the units are no longer observed in the STFT, since then the STFT lacks the temporal resolution needed to distinguish the pauses. Finally as the FFT size increases the number of frequency bins grows, leading to a greater overall computational burden.

The spectrograms of the three simulated image~~s~~ sources at receiver 1 are shown in Fig 3 and the mixtures at each receiver are shown in Fig 4. All the spectrograms shown in this paper were computed using Hanning window, with window size of 1024 and 75% overlap. The results are represented in Fig 5 by the spectrograms of the separated signals as measured at receiver 1. It can be clearly seen that the song units from individual sources are recovered when compared the original images sources with the separated sources. The SDR obtained by the Sawada method in this example is 8.3 dB compared to that achieved by the ideal binary mask which is 12.2 dB.

We also consider the scenario in which the whales are rather more distant from the hydrophone array, specifically $r_1 = 3000$ m, $r_2 = 5000$ m, $r_3 = 4000$ m. In this case the bearings are the same as those used in the section III A. The spectrograms of separated signal are presented in Fig 6 in this case and the SDR which can be achieved is reduced to 4.1 dB, and the lack of separation is evident in the spectrograms. The severe degradation of separation performance is mainly caused by extremely long reverberation time (more than 1.5 s). Specifically, these long reverberation times require a large FFT size in STFT, which

FIG. 3. Spectrograms of the three simulated images sources at receiver x_1, displayed on a normalised dB color scale.

FIG. 4. Spectrograms of the mixtures in two receivers x_1 and x_2, displayed on a normalised dB color scale.

results in the sparseness property being severely degraded.

IV. SOURCE SEPARATION USING REAL RECORDINGS

A. The configuration of hydrophone arrays and the recording system

Our field site is based in the St Marie channel, Madagascar (S16°54′,E49°48′). Humpback whales migrate to the breeding grounds, which include this channel, from Antarctic, and they are present there from July to September. Three hydrophone arrays were employed to record the humpback songs along the coast in 2013. Each array consists of three hydrophones. At each of the three sites there is a central sealed unit which provides amplification and communicates data on-shore. The distance between the centre of the array and each hydrophone is 150 m. The data were transferred through a 1 km cable to the shore. The main power supply for the whole system was located near the coast. The sampling frequency of hydrophone data is 48 kHz, and it is recorded with 16 bit resolution. The hydrophones were placed close to the seabed in a water depth of around 30 m. The hydrophone arrays were recording throughout the breeding period.

For this analysis the recordings were down-sampled to 8 kHz, and the sections containing high source energy and potential multiple sources were identified. In each trial, 60 s song mixtures were separated. Since humpback whales generally do not swim during singing [6], the only source movement is a result of the animals drifting with current, so the whales

FIG. 5. Spectrograms of the separated image sources at receiver x_1, displayed on a normalised dB color scale.

FIG. 6. Spectrograms of the separated image sources at receiver x_1 when the whales are more distant from the hydrophone array.

remain at approximately the same location during this 60 s period.

In such real-world situations the original source signals are unavailable, so one cannot evaluate the separation performance: the SDR metric requires knowledge of the original sources, so is only applicable in simulation, or controlled environments. The separation performance is assessed subjectively by manually listening to the separated songs and visual inspection of the separated song structures observed in the spectrogram. One measure of success of the separation process in real recordings is to look at the structure of the song: if the units are well defined, with clear intervals (pauses) between them then this is suggestive of good separation.

B. Noise reduction based on median threshold scheme

The real recordings are severely contaminated by background noise, including, the noise from the snapping shrimp on the nearby reefs. In conditions when there is significant noise contamination, it is difficult to cluster the observed samples into the original sources in the bin-wise clustering stage. The line orientations cannot be clearly identified because the noise samples are widely distributed over the data space. Assuming a reasonable SNR, the magnitude of the vector $\mathbf{x}(\tau, f)$ is small in periods when the signal is dominated by the noise and large at times when the whale is vocalising. This distinction is lost when the vectors are normalised to have unit norm, as expressed in (7) and as implemented in the standard formulation of the Sawada algorithm. The consequence is that the clustering step is applied to a data for which the signal and noise are not readily distinguished, with the result that the performance of this step can be severely degraded. One primitive approach to reduce the influence of the noise samples in the clustering stage is to remove them, since the noise disturbs the identification of the line orientations.

We propose to discard the observation samples which mainly contain noise in each frequency, and the remaining samples are employed for source separation. The noise reduction is based on the energy of each sample. Specifically, the samples with energy less than the weighted median of all observations in each frequency bin are considered as noise. The median operation is used here, because it is more robust (less sensitive to outliers) compared to mean operation [31]. Let us define a quantity $x_{Med}(f)$ which is the median of observation energy $\sum_j |x_j(\tau, f)|^2$,

$$x_{Med}(f) = Median_\tau(\sum_j |x_j(\tau, f)|^2), \tag{21}$$

where $Median_\tau$ denotes Median operation taken over time. The noise reduction mask based on weighted median $M_{Med}(\tau, f)$ can be constructed as

$$M_{Med}(\tau, f) = \begin{cases} 1, & \text{if } \sum_j |x_j(\tau, f)|^2 \geqslant \beta x_{Med}(f) \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

The threshold for the noise samples in this implementation is determined by a weighting factor β and the median of the time series $x_{Med}(f)$ at frequency f. The parameter β is shared with all frequency bins. The optimum value of β is the one which achieves the largest SDR which can only be determined in simulation scenarios.

The noise reduction is realised by applying the masks $M_{Med}(\tau, f)$ to the original observations,

$$x_j^s(\tau, f) = M_{Med}(\tau, f) x_j(\tau, f). \tag{23}$$

The non-zero samples $x_j^s(\tau, f)$ are utilized in the Sawada method. In order to demonstrate the effectiveness of the noise reduction scheme based on the weighted median threshold, the short range mixtures of artificial humpback song described in Section III A are utilized. Three sources s_1, s_2, s_3 and two receives x_1, x_2 were used in this test. $r_1 = 300$ m, $r_2 = 500$ m, $r_3 = 400$ m. The FFT size is 4096 and the window hop size is 1024. The mixture at each receiver is corrupted by additive Gaussian noise which is filtered so that its spectrum is typical of that ocean ambient noise, with sea state 3 and moderate shipping [32].

Fig 7 compares the performance, in terms of the SDR, for the Sawada algorithm, with and without noise reduction, for various SNR on the input mixture signals. The noise reduction is based on the weighted median threshold and the weighting, $\beta = 1.5$ is used throughout. The separation performance improves significantly through adopting the noise reduction scheme when the SNR is low, which clearly demonstrates the effectiveness of the proposed noise reduction combined with the source separation method. The separation performance

FIG. 7. Comparison of the separation performance for the noisy source separation using the Sawada method with and without noise reduction. The dash line denotes source separation using the Sawada method without noise reduction, whereas, the solid line shows the results when noise reduction is included.

FIG. 8. The spectrograms of the mixtures in three receivers of real recording. Each mixture contains various song units overlapped in the time domain.

drops a little when using the noise reduction in the high SNR regime. This is because some of the source information is misclassified as noise.

The proposed algorithm based on the thresholding scheme within Sawada's method is applied to a field recording of humpback whale song. When employing the algorithm without noise reduction, the separation performance is unreliable and many of the mixtures cannot be separated. However, when including the proposed noise reduction with the Sawada method, most of the mixtures can be successfully separated.

One successful separation example of real recording is detailed in the following. When performing the noise reduction, the parameter β was set as 1.5. The spectrograms of three hydrophone signals collected from the array in St Marie are shown in Fig 8 and the separation results are shown in Fig 9. The bandwidth of spectrograms shown in this example is up to 2 kHz, since the recordings were dominated by noise above 2 kHz. The sampling frequency of the signal is still 8 kHz. The FFT and window hop sizes are 4096 and 1024, respectively, corresponding to a frequency resolution of 2 Hz and time increment of 125 ms. When listening to the mixtures, it is difficult to reliably identify the number of sources. However, we assumed that the number of sources was 3.

The separated sources, shown in Fig 9, are regarded as well separated based on the structure pattern of separated songs. Specifically, the separated source 1 contains mainly loud descending sweeps. The separated source 2 contains mainly sustained tones with harmonic structure. The separated source 3 contains some faint grunts at very low frequency (below 200 Hz), which are quite hard to identify. The regular structures observed in these separated recordings are typical of clean recording on the spectrograms of humpback whale song and so are indicative of successful separation.

FIG. 9. The spectrograms of the separated sources using the noise reduction Sawada method. The separated source 1 contains mainly downward sweeps, while the separated source 2 mainly contains harmonic sustained tones, and the separated source 3 contains faint grunts at very low frequency (below 200 Hz).

V. CONCLUSION

The effectiveness and robustness of the Sawada method for humpback songs separation are verified through separation of artificial mixtures of humpback whale song generated by modelling the underwater impulse responses. For practical implementation on recorded data on an array of 3 hydrophones, the separation performance of the conventional algorithm was largely unacceptable. The primary cause of the poor performance was the high level of noise encountered. This proposed noise reduction method offers significant qualitative improvement in the outputs. The separated signals generated by this approach lack the pristine quality of recordings made near a single singing whale. But do allow one to identify which components in a set of recordings are common to one source.

This paper has demonstrated there is some potential for applying blind source separation method in bioacoustics applications. The potential for success is based on the observations that these vocalisations are usually sparse in the STFT domain, one does, with current methods, need to assume that sources are stationary during the measurement interval. The problems faced in applying the approach include the long reverberation times common in underwater acoustic environment, coupled to high levels of noise. This paper has demonstrated the effectiveness of including noise reduction within the Sawada method and shown that real acoustic environments need not prevent successful separation.

There do remain outstanding problems. Foremost of these is determining the number of sources in practical problems. Further the methods could find more utility if they could be redesigned to operate in an iterative, rather than block-based fashion, which would make them more compatible with real-time implementation.

ACKNOWLEDGMENTS

The authors would like to thank Olivier Adam from University of Paris-Sud for constructive discussions and sharing humpback whale song data for analysis in this paper. They would like to thank staff in CetaMada for providing humpback whale song recordings, and their great support and assistant during the field work in Madagascar. They would also like to thank University of Southampton and CSC (China Scholarship Council) for funding the work.

[1] R. Charif, P. Clapham, and C. Clark, (2001). Acoustic detections of singing humpback whales in deep waters off the British isles, Marine Mammal Science **17**(4), 751-768.

[2] A. K. Stimpert, W. W. L. Au, S. E. Parks, T. Hurst, and D. N. Wiley, (2011). Common humpback whale (*Megaptera novaeangliae*) sound types for passive acoustic monitoring, J. Acoust. Soc. Am. **129**(1), 476-482.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, (2004). Independent component analysis, John Wiley & Sons 1-476.

[4] Y. Razafindrakoto, H. C. Rosenbaum, and D. A. Helweg, (2001). First description of humpback whale song from Antongil Bay, Madagascar, Marine mammal science. **17**(1), 180-186.

[5] R. Suzuki, J. R. Buck, and P. L. Tyack, (2006). Information entropy of humpback whale songs, J. Acoust. Soc. Am. **119**(3), 1849-1866.

[6] E. Mercado, J. N. Schneider, A. A. Pack, and L. M. Herman, (2010). Sound production by singing humpback whales, J. Acoust. Soc. Am. **127**(4), 2678-2691.

[7] E. Mercado, and S. Handel, (2012). Understanding the structure of humpback whale songs

(L), J. Acoust. Soc. Am. **132**(5), 2947-2950.

[8] W. W. L. Au, A. A. Pack, M. O. Lammers, L. M. Herman, M. H. Deakos, and K. Andrews, (2006). Acoustic properties of humpback whale songs, J. Acoust. Soc. Am. **120**(2), 1103-1110.

[9] R. S. Payne and S. McVay, (1971). Songs of humpback whales, Science **173**, 585-597.

[10] E. III. Mercado and L.N. Frazer, (2001). Humpback whale song or humpback whale sonar? A reply to Au et al, IEEE Journal of Oceanic Engineering **26**(3), 406-415.

[11] P. Rickwood, and A. Taylor, (2008). Methods for automatically analyzing humpback song units, J. Acoust. Soc. Am. **123**(3), 1763-1772.

[12] T. A. Abbot, V. E. Premus, and P. A. Abbot, (2010). A real-time method for autonomous passive acoustic detection-classification of humpback whales, J. Acoust. Soc. Am. **127**(5), 2894-2903.

[13] F. Pace, F. Benard, H. Glotin, O. Adam and P. R. White, (2010). Subunit definition and analysis for humpback whale call classification, Applied Acoustics **71**(11), 1107-1112.

[14] L. M. Munger, M. O. Lammers, P. Fisher-Pool, and K. Wong, (2012). Humpback whale (*Megaptera novaeangliae*) song occurrence at American Samoa in long-term passive acoustic recordings, 2008-2009, J. Acoust. Soc. Am. **132**(4), 2265-2272.

[15] M. B. Gur, and C. Niezrecki, (2009). A source separation approach to enhancing marine mammal vocalizations, J. Acoust. Soc. Am. **126**(6), 3062-3070.

[16] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P.Bofill, H.Sawada, A. Ozerov, V. Gowreesunker, D. Lutter and N. Q. K. Duong, (2012). The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges, Signal Processing **92**(8), 1928-1936.

[17] M. I. Mandel, R. J. Weiss and D. P. W. Ellis, (2010). Model-based expectation maximization source separation and localization, IEEE Trans. Audio, Speech and Language Processing **18**(2), 382-394.

[18] N. Q. K. Duong, E. Vincent and R. Gribonval, (2010). Under-determined reverberant audio source separation using a full-rank spatial covariance model, IEEE Trans. Audio, Speech and Language Processing **18**(7), 1830-1840.

[19] A. Ozerov and C. Févotte, (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, IEEE Trans. Audio, Speech and Language Processing **18**(3), 550-563.

[20] H. Sawada, S. Araki, and S. Makino, (2011). Underdetermined convolutive blind source sep-

aration via frequency bin-wise clustering and permutation alignment, IEEE Trans. Audio, Speech and Language Processing **19**(3), 516-527.

[21] A. P. Dempster, N. M. Laird and D. B. Rubin, (1977). Maximum likelihood from incomplete data via the EM algorithm, Journal of the royal statistical society, series B **39**(1), 1-38.

[22] H. Sawada, S. Araki, R. Mukai and S. Makino, (2011). Grouping Separated Frequency Components by Estimating Propagation Model Parameters in Frequency-Domain Blind Source Separation, IEEE Trans. Audio, Speech and Language Processing **19**(15), 1592-1604.

[23] Ö. Yilmaz and S. Rickard, (2004). Blind separation of speech mixtures via time-frequency masking, IEEE Trans. Signal Processing **52**(7), 1830-1847.

[24] P. D. O'Grady and B. A. Pearlmutter, (2008). The LOST algorithm: finding lines and separating speech mixtures, EURASIP Journal on Advances in Signal Processing **2008**(1), 1-17.

[25] H. Sawada, S. Araki and S. Makino, (2007). Measuring Dependence of Bin-wise Separated Signals for Permutation Alignment in Frequency-domain BSS, ISCAS 3247-3250.

[26] A. L. Magg and A. J. Duncan, Acoustic toolbox user-interface & post-processor, installation & user guide, Curtin University of Technology.

[27] E. L. Hamilton, (1976). Shear-wave velocity versus depth in marine sediments: a review, Journal of Geophysics **39**(41), 985-996.

[28] E. L. Hamilton, (1972). Compressional-wave attenuation in marine sediments, Journal of Geophysics **39**(37), 620-646.

[29] B. A. Brunson and R. K. Johnson, (1980). Laboratory measurements of shear wave attenuation in saturated sand, J. Acoust. Soc. Am. **39**(68), 1371-1375.

[30] E. Vincent, R. Gribonval and C. Févotte, (2006). Performance measurement in blind audio source separation, IEEE Trans. Audio, Speech and Language Processing **14**(4), 1462-1469.

[31] T. S. T. Leung, and P. R. White, (1998). Robust estimation of oceanic background noise spectrum, Mathematics in Signal Processing IV, Clarendon Press, Oxford 369-382.

[32] R. Uric, (1967). Principles of underwater sound for engineers, Chapter 7: the noise background of the sea: ambient-noise level, Tata McGraw-Hill Education.

Fig. 1. The simulation experimental configuration for sources and receivers in the underwater environment. x_1, x_2 represent two receivers, and s_1, s_2, s_3 denote three sources.

Fig. 2. The impulse responses for both the source and the receiver located depth 50 m and their distance apart is 510 m in the left panel and 3131 m in the right panel. The origin of the time axis is defined so that the majority of the propagation delay has been removed.

Fig. 3. Spectrograms of the three simulated images sources at receiver x_1, displayed on a normalised dB color scale.

Fig. 4. Spectrograms of the mixtures in two receivers x_1 and x_2, displayed on a normalised dB color scale.

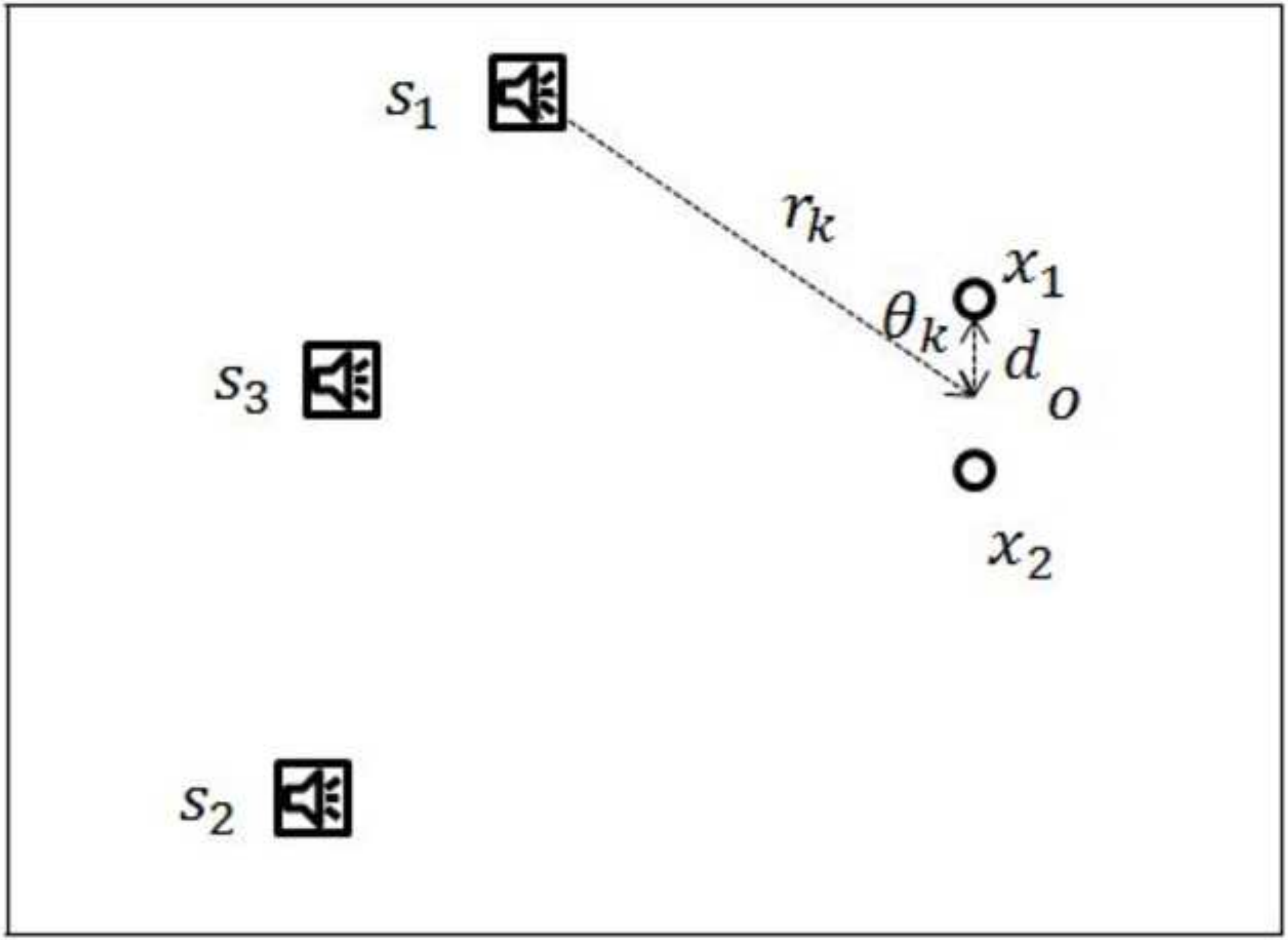
Fig. 5. Spectrograms of the separated image sources at receiver x_1, displayed on a normalised dB color scale.

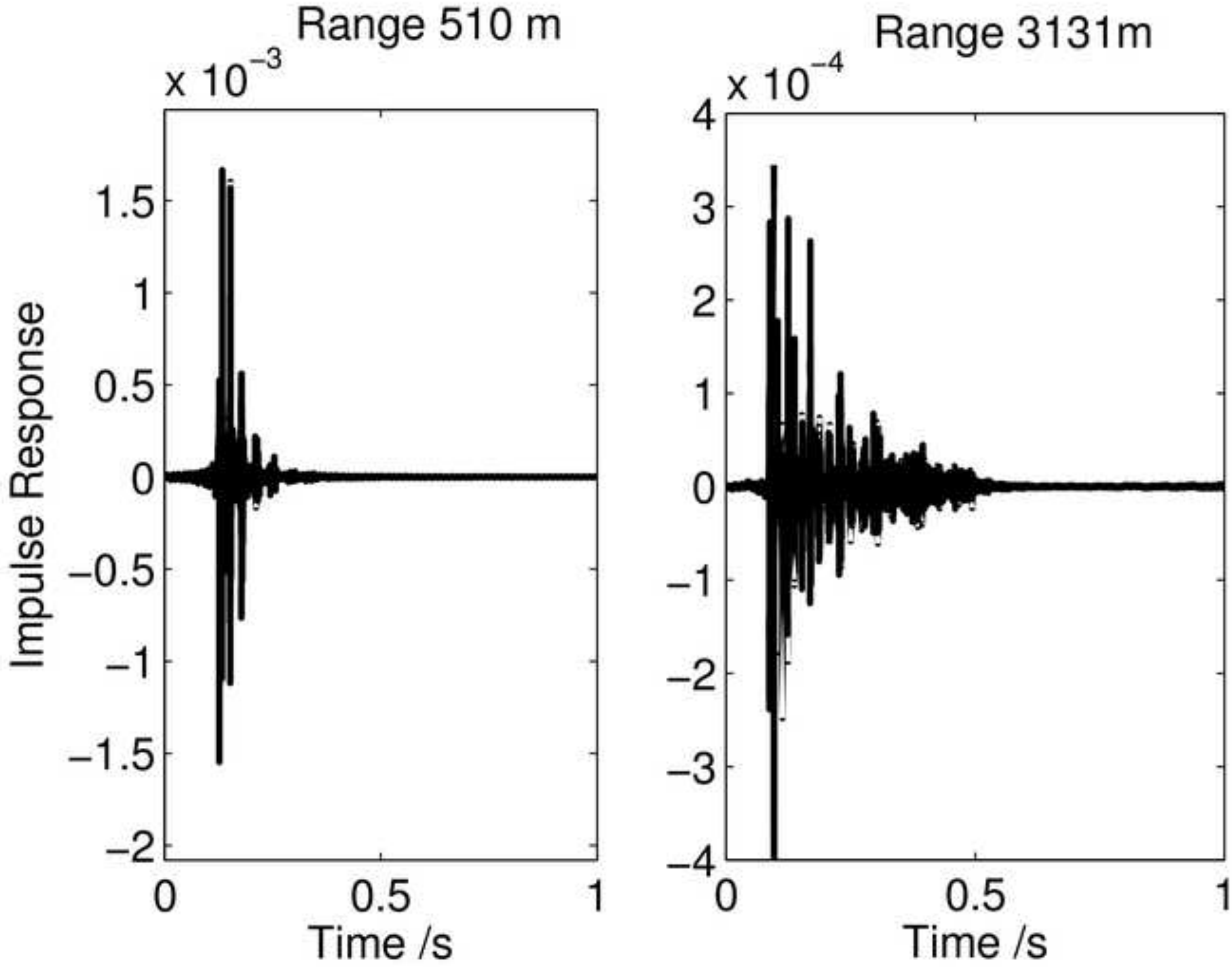
Fig. 6. Spectrograms of the separated image sources at receiver x_1 when the whales are more distant from the hydrophone array.

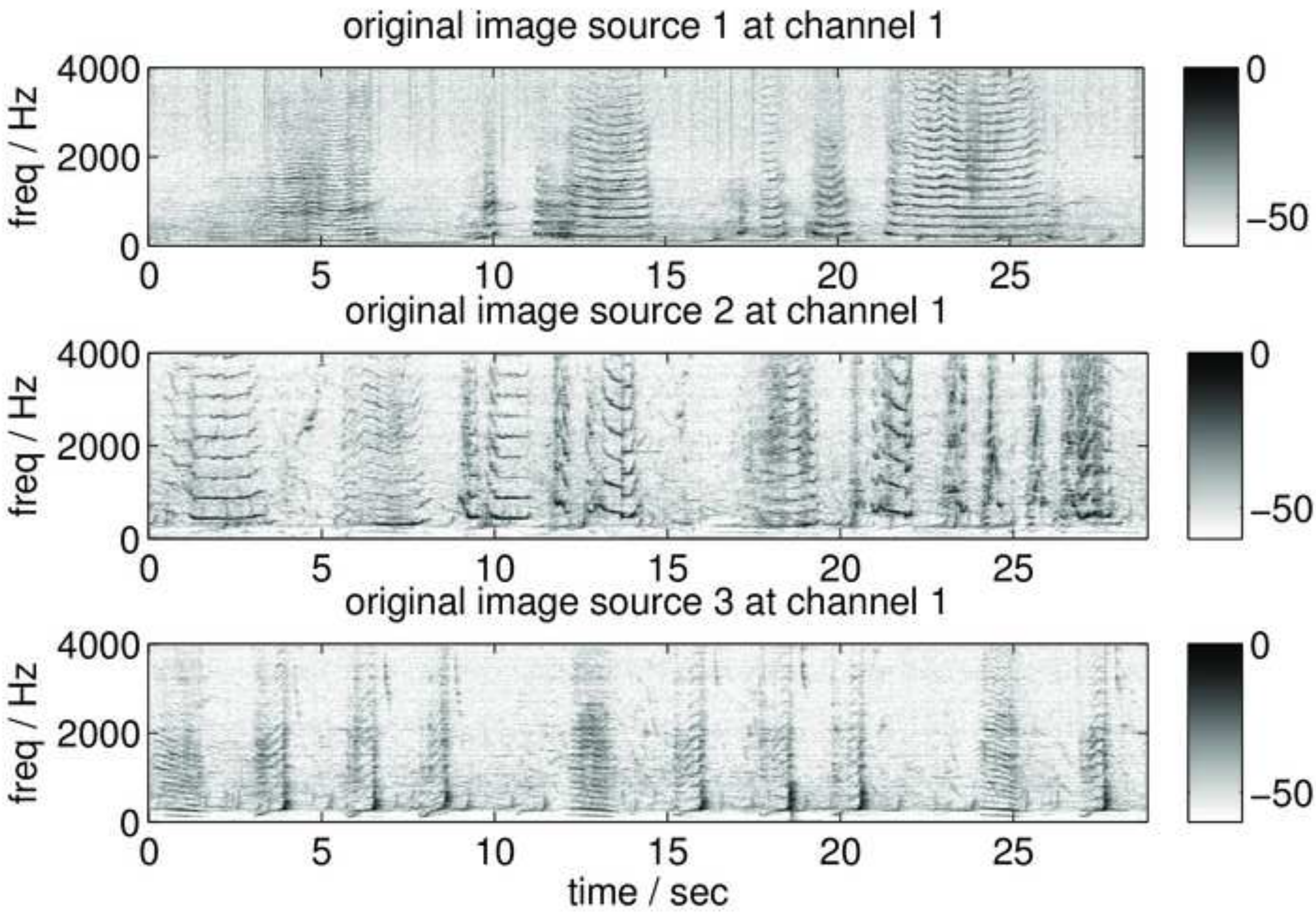
Fig. 7. Comparison of the separation performance for the noisy source separation using the Sawada method with and without noise reduction. The dash line denotes source separation using the Sawada method without noise reduction, whereas, the solid line shows the results when noise reduction is included.

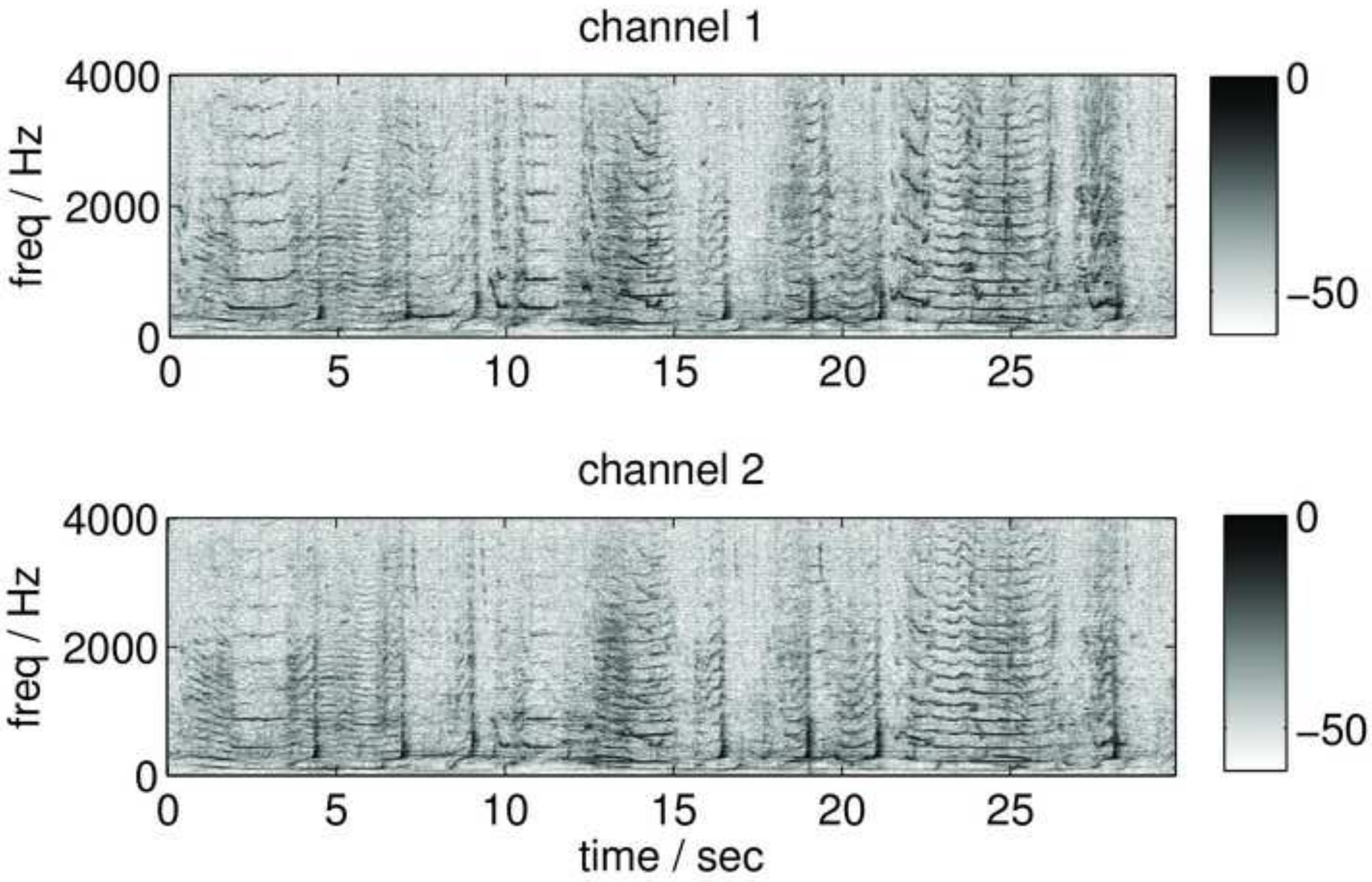
Fig. 8. The spectrograms of the mixtures in three receivers of real recording. Each mixture contains various song units overlapped in the time domain.

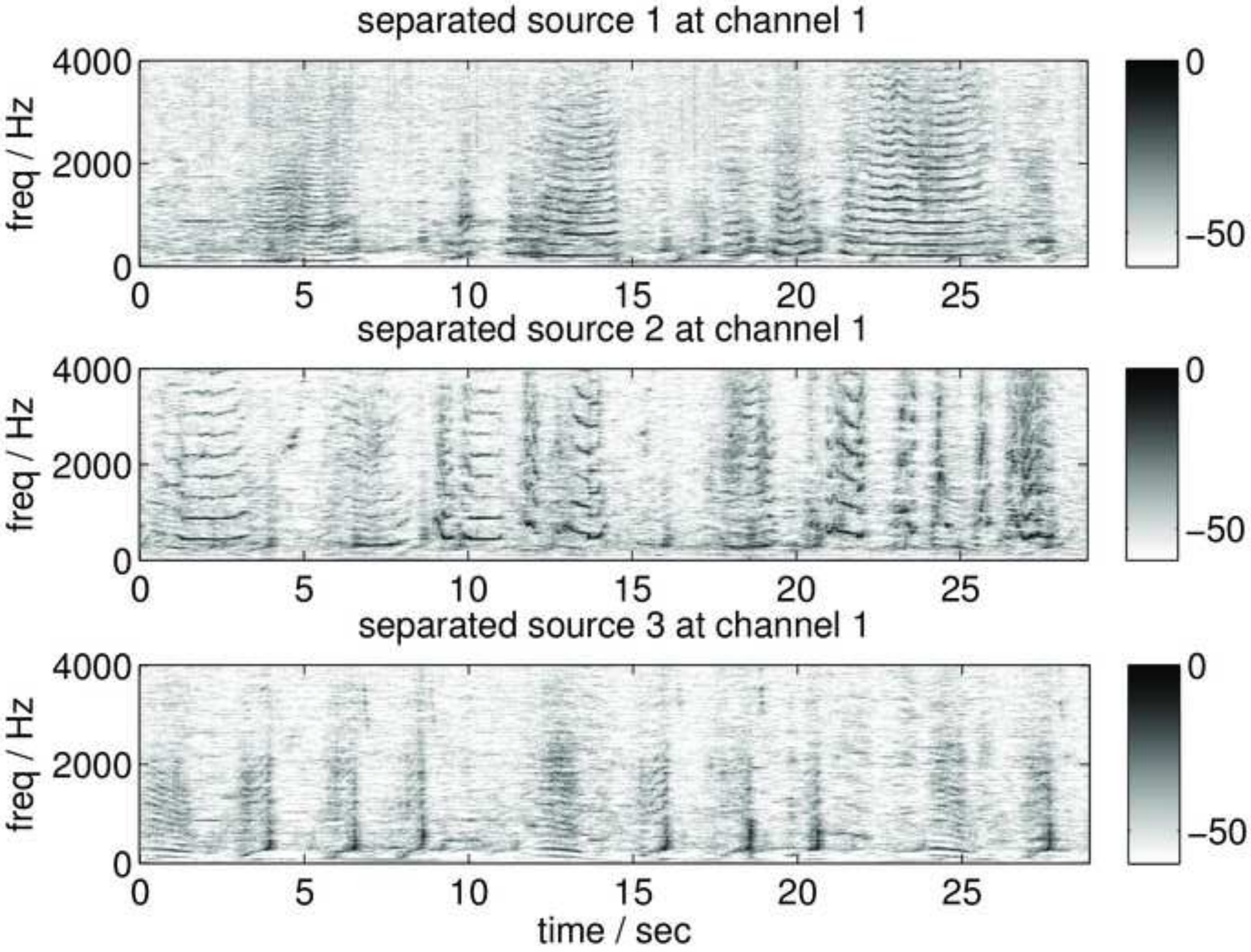
Fig. 9. The spectrograms of the separated sources using the noise reduction Sawada method. The separated source 1 contains mainly downward sweeps, while the separated source 2 mainly contains harmonic sustained tones, and the separated source 3 contains faint grunts at very low frequency (below 200 Hz).


s_1
r_k
x_1
θ_k
d
o
s_3
x_2
s_2

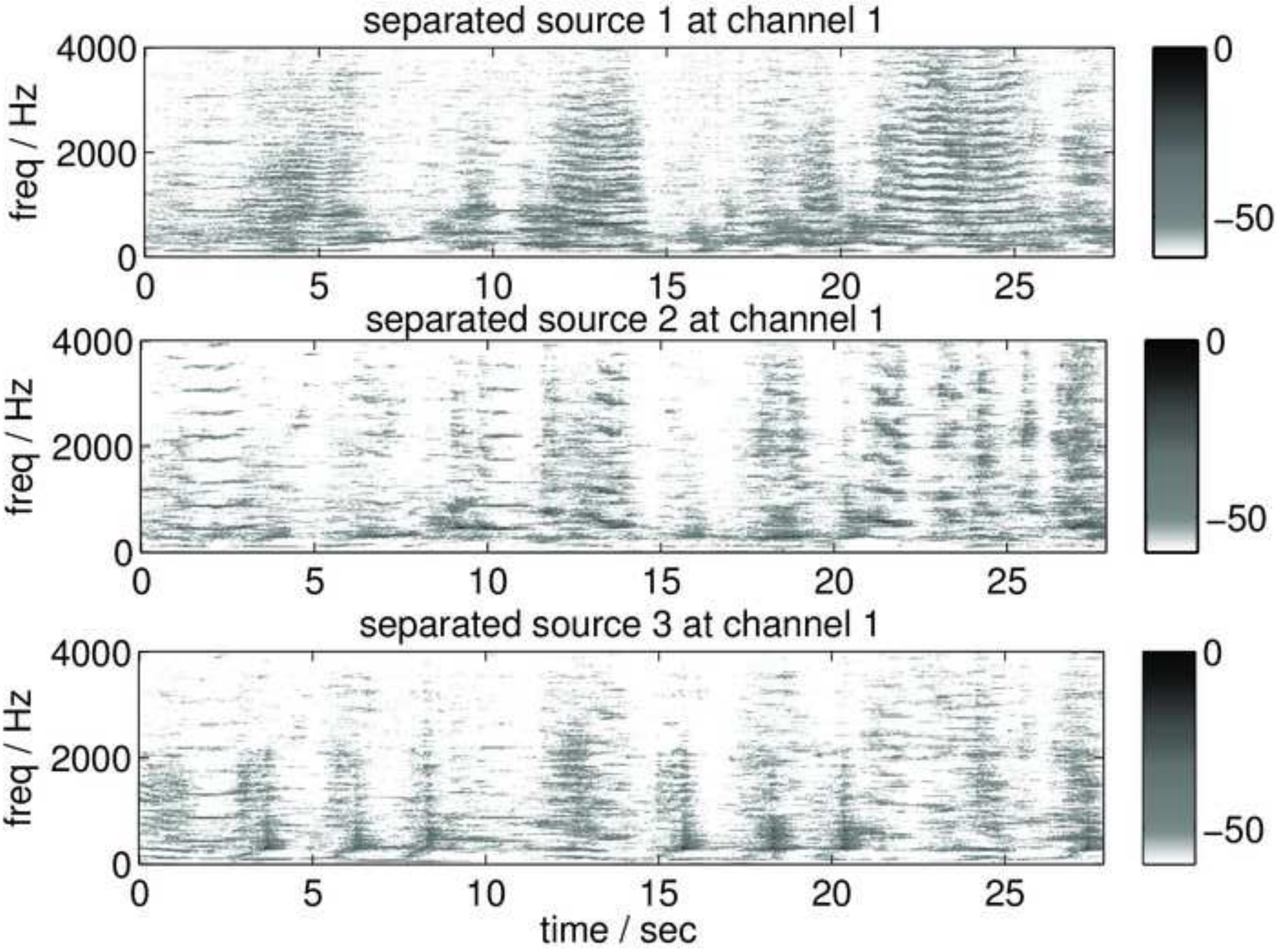

Range 510 m
x 10^-3
1.5
1
0.5
0
−0.5
−1
−1.5
−2
Impulse Response
0
0.5
1
Time /s
Range 3131m
x 10^-4
4
3
2
1
0
−1
−2
−3
−4
0
0.5
1
Time /s

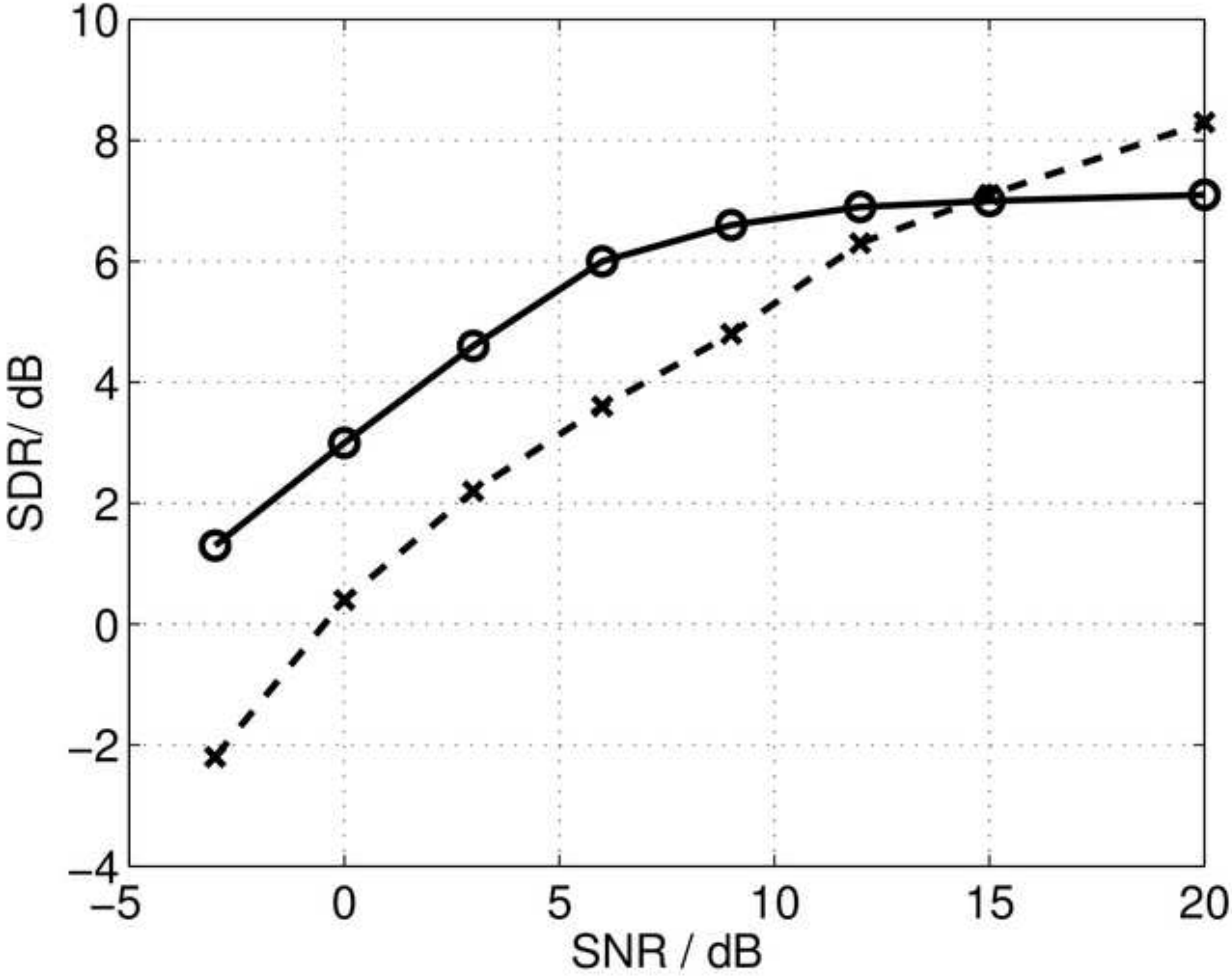

original image source 1 at channel 1
freq / Hz
4000
2000
0
0
5
10
15
20
25
0
−50
original image source 2 at channel 1
freq / Hz
4000
2000
0
0
5
10
15
20
25
0
−50
original image source 3 at channel 1
freq / Hz
4000
2000
0
0
5
10
15
20
25
0
−50
time / sec

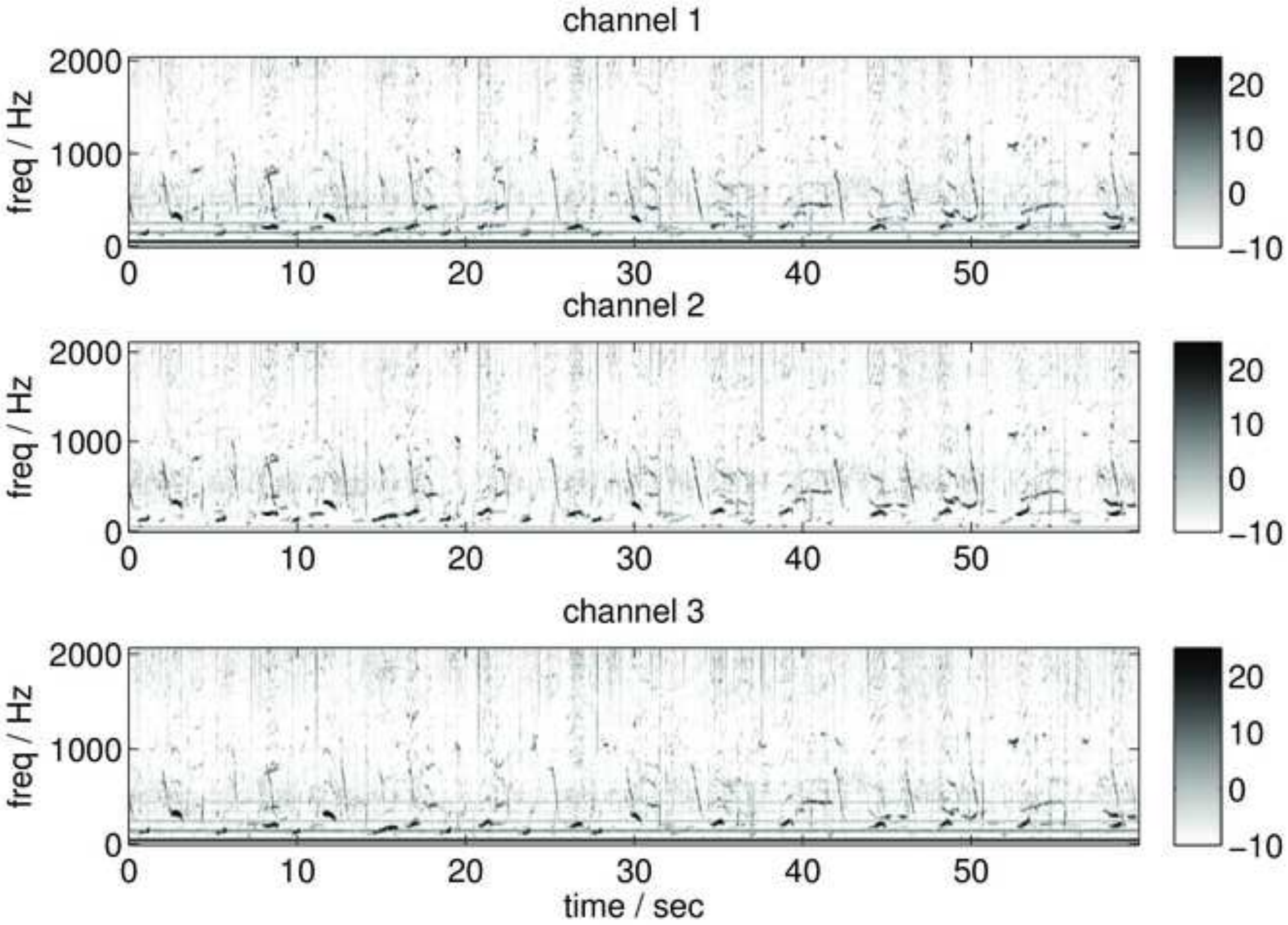

channel 1
4000
2000
0
freq / Hz
0
5
10
15
20
25
0
−50
channel 2
4000
2000
0
freq / Hz
0
5
10
15
20
25
time / sec
0
−50

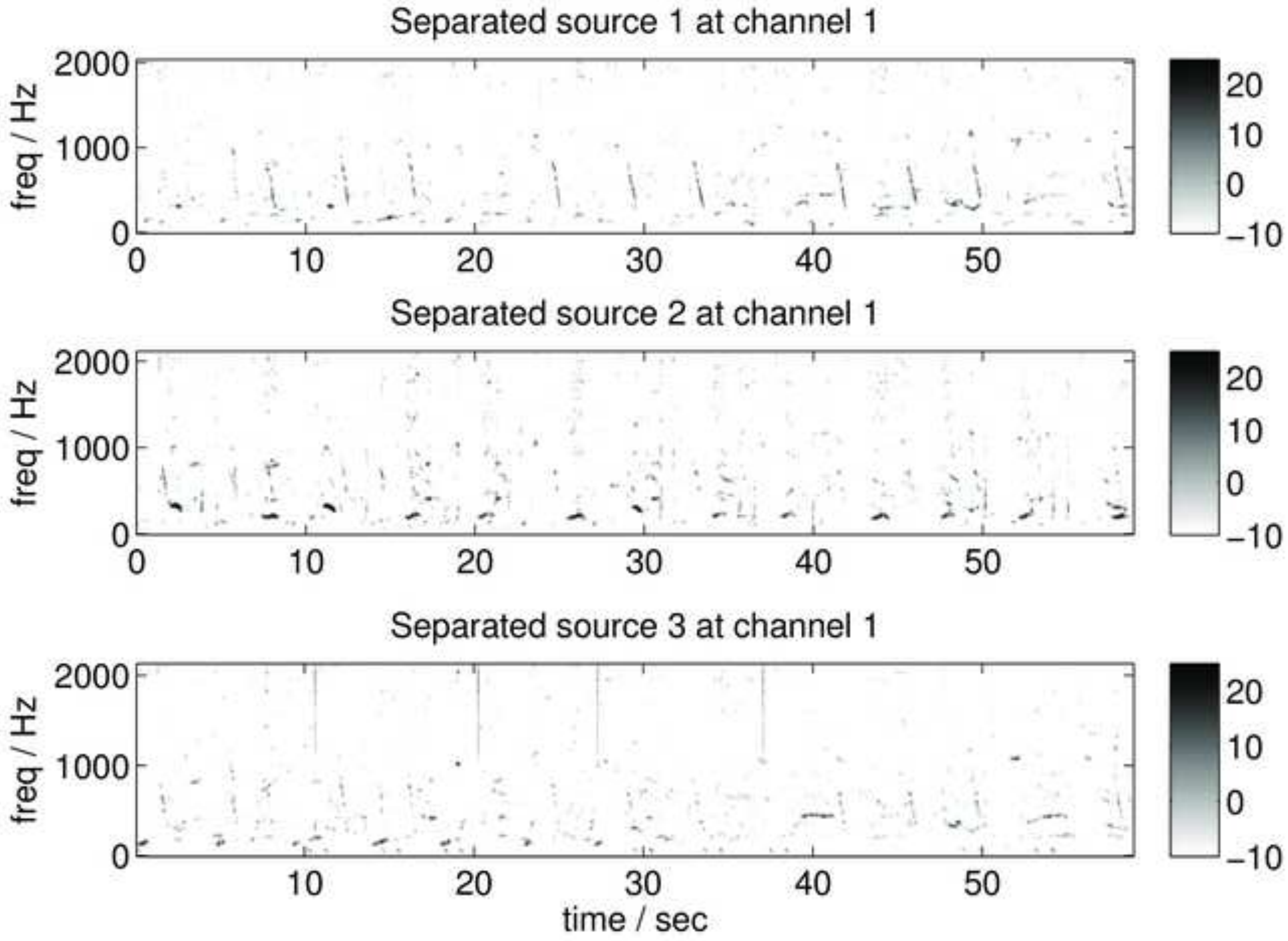

separated source 1 at channel 1
separated source 2 at channel 1
separated source 3 at channel 1
freq / Hz
4000
2000
0
0
5
10
15
20
25
time / sec
0
−50

Click here to download Figures (TIFF, EPS, PS, or JPEG only) Figure6.jpg




10
8
6
4
2
0
−2
−4
−5
0
5
10
15
20
SNR / dB
SDR/ dB


channel 1
freq / Hz
2000
1000
0
0
10
20
30
40
50
20
10
0
−10
channel 2
freq / Hz
2000
1000
0
0
10
20
30
40
50
20
10
0
−10
channel 3
freq / Hz
2000
1000
0
0
10
20
30
40
50
time / sec
20
10
0
−10


Separated source 1 at channel 1
Separated source 2 at channel 1
Separated source 3 at channel 1
freq / Hz
2000
1000
0
0
10
20
30
40
50
time / sec
20
10
0
−10