STATISTICAL PROPERTIES OF STFT RATIOS FOR TWO CHANNEL SYSTEMS AND APPLICATIONS TO BLIND SOURCE SEPARATION

Radu Balan and Justinian Rosca

{rvbalan,rosca}@scr.siemens.com
Siemens Corporate Research
Princeton, NJ 08540

ABSTRACT

The ratio of the short time Fourier transform (STFT) coefficients of signals received at two sensors can factor out the role of the power spectrum of emitting sources, under an assumption called disjoint orthogonality. Thus, it can reveal parameters specific to the mixing scenario and serve as a basis for channel estimation techniques.

In this paper we analyze and extend a source separation method based on the use of STFT ratios of two sensor inputs, called DUET. We generalize the problem formulation in DUET, and prove that considerably weaker assumptions about the classes of input signals are sufficient to apply the derived techniques. The analysis centers on the notion of *ratio-estimator* and a novel stochastic model that enabled us to derive the maximum likelihood ratio-estimator.

Derived techniques can be effectively applied to source separation, source localization, signal enhancement, and noise reduction when using a twin microphone system, both in echoic environments and degenerate situations.[1]

1. INTRODUCTION

Miniaturized sensors and increased computational power and memory storage in today's digital signal processors make it possible to implement and apply advanced DSP techniques to problems of source separation and noise reduction for small electronic devices (e.g. speech recognition front ends, personal digital assistants with voice input, mobile phones, smart alarms etc.). Such devices can take advantage of two or more microphone arrays, and are aimed at improving the directionality of the signal input system, or simply of source separation, while not affecting the quality of the sound (particularly if the sound of interest is speech). In recent years, this domain has been the focus, at the low end of applications, for Blind Source Separation (BSS) and Independent Component Analysis (ICA) techniques [4, 8, 12, 2]. Traditionally though, array processing and beamforming signal processing techniques were concerned with the formation of steered beams for an array of sensors in sonar and radar systems [5, 11].

In this paper we analyze and extend a source separation method based on the idea of interpreting STFT ratios of two sensor inputs, which has been recently proposed in [3] under the name DUET (Degenerate Unmixing and Estimation Technique). We start with the problem formulation in DUET (see Section 2), and prove that weaker assumptions about the classes of input signals combined with an optimal statistical interpretation of the data are sufficient to solve the same problem.

Rather than learn filters to demix data according to some statistical criterion [10], our analysis centers on the idea of estimating ratios of two transfer functions H_{1i}/H_{2i} (see Section 3), and using these estimates to further infer a parameterized model and demix sources. The mixing model we assume is the following:

$$\begin{aligned} x_1(t) = &\quad h_{11} \star s_1(t) + \cdots + h_{1N} \star s_N(t) \\ x_2(t) = &\quad h_{21} \star s_1(t) + \cdots + h_{2N} \star s_N(t) \end{aligned} \tag{1}$$

where $s_1, \ldots, s_N$ are N sources of interest, x_1, x_2 are the sensor measurements and $h_{11}, \ldots, h_{2N}$ are the $2N$ channel formal impulse responses. By *formal* we mean that fractional delays [6] are allowed.

Besides statistical independence of the sources, we make two other basic assumptions:

- Sources are stationary on a short-time horizon, but their frequency content has jumps over long-time periods;
- For a given time window, signals may have frequency gaps, but in the long term they cover all the frequency bands (*ergodicity* or *persistence* hypothesis);

These qualitative properties will form the basis of a mathematical model for a class of signals of interest. Regarding the channel description, we do not make any assumption at this time. However further into our analysis we apply our technique to both anechoic and echoic mixing models. Although most of this paper is concerned with the two microphone case, we also present extensions of the techniques introduced to the multi-sensor case.

The layout of this document is as follows. Next section briefly reviews elements of interest about beamforming and degenerate demixing. Section 3 describes the principles of our statistical approach, a formalization of the estimation problem, and its consequences. Section 4 presents an experimental validation of our approach, and is followed by conclusions and ideas for future work.

2. BACKGROUND

The basic principle for singling out a source by beamforming can be used in adaptive algorithms for demixing real-world

[1] Presented at International Workshop on ICA and BSS 2000.

anechoic and echoic signals [9, 3]. Of particular interest here is the technique for estimating mixing model parameters in [3], which was applied for degenerate mixtures (more sources than two and two microphones). We briefly review these principles below.

2.1. Beamforming

Beamforming is the problem of processing signals impinging on an array of sensors in order to maximize reception from a given direction. The sonar and radar applications of beamforming principles led directly to an early study of the topic and a rich literature [5, 11]. The array response or direction vector can be easily determined for a finite speed of propagation c of a planar wave in a real environment and an equidistant linear array of sensors, spaced by distance d. One source emitting from direction angle θ_i is delayed at the next adjacent sensor by $\delta_i = \frac{d}{c}\cos\theta_i$. Assuming the source is very far compared to d the effect induced in a K-array is:

$$X(\omega) = S(\omega) \cdot \left[\ 1 \quad e^{-j\omega\delta_i} \quad \dots \quad e^{-j\omega(K-1)\delta_i}\ \right] \tag{2}$$

where X, S are the Discrete Fourier Transform of the measurements and signal, respectively. The DFTs can be replaced by short-time Fourier transforms, defined as follows. For a source signal $s(t)$, $t = 0, \dots, L-1$, and a fixed window w, $t = 0, \dots, M-1$ $(M < L)$:

$$S(\omega, k) = \sum_{l=k}^{M+k} e^{-\frac{2\pi j}{M}\omega} w(l - kb)s(l),\ k = 0, 1, \dots, B = \lfloor \frac{L}{M} \rfloor \tag{3}$$

where b (the time step) is usually a fraction of M, the window size. The ratio $\frac{b}{M}$ represents the redundancy of this representation. If $w \equiv 1$ and $b = L = M$, one recovers the usual definition of the DFT (Discrete Fourier Transform). In [1] we analyzed the influence of the window w on the formal manipulations of delays. In essence, we proved that the windowing effect is negligible. As a result, we fix the window w to a particular form (for instance the Hanning window).

A linear combination of the measurements at the sensors defines a spatial filter that improves the reception of a narrow band source. Adaptive techniques can discover and focus on one source at a time. This formulation of the problem resulted in a successful BSS approach to real-world broadband (audio) signals using only two closely spaced microphones [9].

Frequency domain approaches to beamforming problem equally allow the recovery of a source of interest from input mixtures. Although such approaches are considered to be computationally intensive, the challenging part is the simultaneous learning of demixing parameters at all frequencies. There is a strong analogy between our approach and frequency domain beamforming in the principle and architecture for signal processing. However the way in which parameters are learnt is radically different.

2.2. STFT ratios for two channel systems

[3] introduced the DUET technique for blind separation of an arbitrary number of sources from two mixtures of the sources, and claimed that it works under particular assumptions about the sources. In general, the BSS literature makes use of either the **statistical independence** assumption or the **statistical orthogonality** assumption. In contrast, DUET introduced an assumption called *disjoint orthogonality.* By definition, N sources $s_1, s_2, ..., s_N$ are **disjointly orthogonal** iff:

$$S_i(\omega) \cdot S_j(\omega) = 0, \forall\omega, \text{ and } \forall i \neq j \tag{4}$$

where these are the DFT transforms of the signals. This means that at most one source has a nonzero Fourier component for any frequency ω. In practice finite windows of data are used, hence the analogous property for windowed Fourier transforms and a fixed windowing function $w(t)$, called *w-disjoint orthogonality*, is:

$$S_i(\omega, \tau) \cdot S_j(\omega, \tau) = 0, \forall\omega, \tau \text{ and } \forall i \neq j \tag{5}$$

Under the disjoint orthogonality assumption, it is obvious that at most one of the N sources, let it be s_i, will be non-zero for a given frequency ω. In the anechoic model, in the same way as in beamforming theory for plane waves impinging on two sensors, a source emitting from direction angle θ_i will be delayed at the second sensor by $\delta_i = \frac{d}{c}\cos\theta_i$, and will be possibly attenuated by factor a_i. Therefore:

$$\begin{aligned} X_1(\omega) &= S_i(\omega) \\ X_2(\omega) &= S_i(\omega) \cdot a_i \cdot e^{-j\omega\delta_i} \end{aligned} \tag{6}$$

The i^{th} source's parameters a_i and δ_i can be obtained as follows from these relations:

$$a_i = \left| \frac{X_2(\omega)}{X_1(\omega)} \right|, \quad \delta_i = \frac{1}{\omega}\mathrm{Im}(\ln\frac{X_1(\omega)}{X_2(\omega)}) \tag{7}$$

The ratio $\frac{X_1(\omega)}{X_2(\omega)}$ is an *STFT ratio* and the parameters derived from it is what we call *ratio-estimates.* In theory, DUET can determine all ratio-estimates (parameters a_i and δ_i in this case) by detecting N peaks of clusters in an amplitude-delay histogram defined by the equations above. Then it can obtain estimates of the sources from one mixture only by selecting the corresponding frequencies and transforming back to the time domain.

In practice, the frequencies corresponding to one cluster Ω_i do not result exactly in the same δ_i. To account for noise and estimation errors, DUET uses an averaging estimate:

$$\hat{\delta}_i = \frac{1}{|\Omega_i|}\sum_{\omega\in\Omega_i}\frac{1}{\omega}\mathrm{Im}(\ln\frac{X_1(\omega)}{X_2(\omega)}) \tag{8}$$

Even so, DUET can demix surprisingly well a mixture of five speech sources, but estimated sources have serious artifacts. Artifacts increase as the number of sources increases. The disjoint orthogonality assumption may be too strong a condition, which is not satisfied in reality. For one thing, disjoint orthogonality implies statistical orthogonality. Many successful BSS approaches rely, in the $N \times N$ case, simply on statistical orthogonality [7]. Is the disjoint orthogonality assumption really necessary? The assumption becomes unrealistic especially when more than two sources are mixed together.

We show that the disjoint orthogonality assumption is not necessary when applying ratio-estimates of the form given in Equation 7. Mixing model parameters can be estimated using statistical techniques under much broader conditions, which we define next.

3. STATISTICAL APPROACH

We generalize ratio-estimates, formally introduce the class of signals for which statistical properties of ratio-estimates are relevant, and derive an optimal ratio-estimator. We claim that ratio-estimates facilitate solutions to the types of BSS applications mentioned.

3.1. Stochastic Model for Signals of Interest

Let us consider again the convolutive mixing model (1). The transfer functions to the second microphone can be included in the source definitions. After redefinition:

$$\begin{aligned} x_1(t) = &\quad r_{11} \star s_1(t) + \cdots + r_{2N} \star s_N(t) \\ x_2(t) = &\quad s_1(t) + \cdots + s_N(t) \end{aligned} \tag{9}$$

In the frequency domain, the above equations become:

$$\begin{aligned} X_1(\omega, k) &= R_{11}(\omega) S_1(\omega, k) + \cdots + R_{1N}(\omega) S_N(\omega, k) \\ X_2(\omega, k) &= S_1(\omega, k) + \cdots + (\omega) S_N(\omega, k) \end{aligned} \tag{10}$$

where $S_1, \ldots, S_N, X_1, X_2$ are the source and measurement short-time Fourier transforms. Because windowing effects are negligible [1], the unknown R coefficients are the same as in the case of the regular Fourier transform, for all practical purposes. R plays the role of a ratio of transfer functions. Our problem is to estimate R.

Rather than simplify the expressions above using the disjoint orthogonality assumption introduced in the previous section, we interpret them from a statistical perspective. In general, sources can use simultaneously the same frequency ω_0. However, many frequencies are available so that there must be cases when sources do not use ω_0 over a short period of time. A statistical analysis of the STFT ratio $\frac{X_1(\omega)}{X_2(\omega)}$ should separate the situation when one and only one source uses a particular frequency from the case when none or all use ω_0. The former case enables us to reliably estimate the parameters of the source propagation model from data, and therefore separate sources in the end. Indeed for those cases when only one source emits at ω_0, $\frac{X_1(\omega_0)}{X_2(\omega_0)}$ reduces to one of $R_{1i}(\omega_0)$ for some i. At this point a model of source propagation can be used, and its parameters can be estimated or the statistical properties of the ratio can be directly used.

Our question then is: Is it possible at all to estimate R values reliably given that noise is present and the disjoint orthogonality assumption does not necessarily hold?

At this point we recall the qualitative statistical assumptions made in Section 1. Sources should be independent, stationary over short-time periods but with discontinuous spectral content over long-time intervals, and be ergodic (or persistent). Now, we can make these hypotheses more precise. We construct a stochastic model that satisfies all these requirements and naturally relaxes the strong w-disjoint orthogonality assumption used in the DUET algorithm.

We recast short-term stationarity into an assumption of independence of the short-time frequency components. Thus, for every fixed k, $S(\omega_1, k)$ is independent of $S(\omega_2, k)$ for $\omega_1 \neq \omega_2$. Note that this would be formally true if samples were jointly Gaussian. Next, we model the discontinuous behavior of the spectral power as a product of two random variables: a continuous (or quasi-continuous, i.e. a discretized continuous) random variable, denoted by G, and a Bernoulli random variable V with probability p of being 1 and $(1-p)$ of being 0. This is the crucial assumption of our model. In Section 4 we present experimental evidence to support our stochastic model. The intuition behind it comes from the time-frequency representation of the speech. In the time-frequency (TF) plane, speech forms various ridge patterns. Consider that for a fixed frequency, one sees a nonzero spectral power on that frequency channel for a given time-frame. The energy pulse may go on into the next time frame, branch into adjacent frequency bands, or simply disappear. Such a behavior suggests modeling the TF components $S(\omega, k)$ as a product of two random variables (RV):

$$S(\omega, k) = V^{\omega,k} G^{\omega,k} \tag{11}$$

where V is a Bernoulli RV or switching process and G is a continous RV. For the purpose of this paper we consider that the spectral components are independent for different time frame indices. Thus, in fact, we assume $S(\omega_1, k_1)$ is independent of $S(\omega_2, k_2)$ for every $(\omega_1, k_1) \neq (\omega_2, k_2)$. For speech signals this hypothesis can be relaxed to accommodate, for instance, a hidden Markov model. Finally the ergodicity (or persistence) hypothesis allows us to assume that for every frequency ω, $V(\omega)$ is of non-vanishing variance. This assumption is by no means essential to our algorithms, and in fact in several applications we tune the frequency set on a particular signal of interest. However we make this hypothesis to avoid degenerate cases. We summarize our stochastic model below:

Signal Class. The class of signals of interest is formed by those stochastic signals whose short-time Fourier transform is factorized as a product of a discrete Bernoulli RV and a continuous (or quasi-continuous) RV as in (11).

3.2. The Optimal ML Ratio-Estimator

The main goal here is to define an optimal estimator for R using the stochastic model introduced before. Consider the two-source case for which the mixing model (10) turns into:

$$\begin{aligned} X_1(\omega, k) = &\quad R_1(\omega) V_1^{\omega,k} G_1^{\omega,k} + R_2(\omega) V_2^{\omega,k} G_2^{\omega,k} \\ X_2(\omega, k) = &\quad V_1^{\omega,k} G_1^{\omega,k} + V_2^{\omega,k} G_2^{\omega,k}, \qquad k = 1, \ldots, B \end{aligned} \tag{12}$$

For the remainder of this subsection we fix a frequency ω and we omit writing it. For this model we make the following assumptions: (1) V_1, V_2 are Bernoulli random variables with probabilities of success p_1, p_2, respectively; (2) G_1, G_2 are discrete random variables, uniformly distributed over a sufficiently large set of equispaced points (say K_1, K_2); (3) $V_1^k, V_2^k, G_1^k, G_2^k$ are i.i.d. copies of the random variables V_i, G_i .

Our problem is to estimate R_1 and R_2 based on B block data measurements $X_1(1), \ldots, X_1(B)$ and $X_2(1), \ldots, X_2(B)$.

We compute the maximum likelihood estimator for R_1, R_2 by conditioning with respect to V_1^k, V_2^k. At every block k:

$$Pr(X_1(k), X_2(k)|R_1, R_2) =$$
$$\sum_{a,b\in\{0,1\}} Pr(X_1(k), X_2(k)|R_1, R_2, V_1^k = a, V_2^k = b) Pr_{V_1}(a) Pr_{V_2}(b)$$

Thus for the likelihood we obtain:

$$\Pr(X_1, X_2|R_1, R_2) =$$
$$\prod_{k=1}^{B}[(1-p_1)(1-p_2)\cdot \Pr(X_1(k)=0, X_2(k)=0) +$$
$$+ p_1(1-p_2)\cdot \Pr(X_1(k) = R_1 G_1^k, X_2(k) = G_1^k|R_1) +$$
$$+ (1-p_1)p_2 \cdot \Pr(X_1(k) = R_2 G_2^k, X_2(k) = G_2^k|R_2) +$$
$$+ p_1 p_2 \cdot \Pr(X_1(k) = R_1 G_1^k + R_2 G_2^k, X_2(k) = G_1^k + G_2^k|R_1, R_2)]$$

where X_1, X_2 are B-vectors of complex numbers. Note that the middle term probabilities can be written as:

$$\Pr(X_1(k) = R_1 G_1^k, X_2(k) = G_1^k|R_1) =$$
$$= \delta(X_1(k) = R_1 X_2(k)) \cdot P_{G_1}(X_2(k))$$

$$\Pr(X_1(k) = R_2 G_2^k, X_2(k) = G_2^k|R_2) =$$
$$= \delta(X_1(k) = R_2 X_2(k)) \cdot P_{G_2}(X_2(k))$$

where $\delta()$ is the Kronecker symbol. Note that X_1, X_2 can take values only on a discrete lattice. The lattice structure has the following consequences:

Lemma 1. For nonzero $R_1 \neq R_2$, the following implications hold true for the events (a),(b), and (c) defined below:

$$\begin{array}{ll}
(a) & X_1(k) = 0 \,\&\, X_2(k) = 0 \\
 & \Rightarrow X_1(k) = R_1 X_2(k) \,\&\, X_1(k) = R_2 X_2(k) \\
(b) & (X_1(k) \neq 0 \text{ or } X_2(k) \neq 0) \,\&\, X_1(k) = R_1 X_2(k) \\
 & \Rightarrow X_1(k) \neq R_2 X_2(k) \\
(c) & (X_1(k) \neq 0 \text{ or } X_2(k) \neq 0) \,\&\, X_1(k) = R_2 X_2(k) \\
 & \Rightarrow X_1(k) \neq R_1 X_2(k)
\end{array}$$

Lemma 2. Events (a),(b), and (c) are mutually exclusive.

Lemma 3. The set $\{1, \ldots, B\}$ can be disjointly partitioned into four sets T_1, T_2, T_3, T_4 as determined by events (a),(b), and (c).

$$\begin{array}{ll}
T1 = & \{i \in \{1, \ldots, B\} | (a) \text{ is true}\} \\
T2 = & \{i \in \{1, \ldots, B\} | (b) \text{ is true}\} \\
T3 = & \{i \in \{1, \ldots, B\} | (c) \text{ is true}\} \\
T4 = & \{i \in \{1, \ldots, B\} | (a) \text{ and } (b) \text{ and } (c) \text{ are false}\}
\end{array}$$

Now, we can prove our main result:

Theorem. For uniformly distributed G_i the likelihood $\Pr(X_1, X_2|R_1, R_2)$ has the form:

$$\Pr(X_1, X_2|R_1, R_2) =$$
$$[(1-p_1)(1-p_2) + \frac{p_1(1-p_2)}{K_1} + \frac{p_2(1-p_1)}{K_2} + \frac{p_1 p_2}{K_1 K_2}]^{|T_1|}$$
$$\cdot [\frac{p_1(1-p_2)}{K_1} + \frac{p_1 p_2}{K_1 K_2}]^{|T_2|} \cdot [\frac{p_2(1-p_1)}{K_2} + \frac{p_1 p_2}{K_1 K_2}]^{|T_3|} \cdot [\frac{p_1 p_2}{K_1 K_2}]^{|T_4|}$$
$$= [\frac{p_1 p_2}{K_1 K_2}]^B \cdot [1+E_1]^{|T_1|} \cdot [1+E_2]^{|T_2|} \cdot [1+E_3]^{|T_3|}$$

where $|T_l|$ denotes the cardinality of the set T_l.

Proof. In the expression of $\Pr(X_1, X_2|R_1, R_2)$, the product $\prod_{k=1}^{B}$ is split into four sub-products according to which one of the four sets T (Corrolary 2) i belongs.

Note that E_1, E_2, E_3 and $|T_1|$ do not depend on R_1, R_2, but rather on the prior information. Also p_1, p_2, K_1, K_2 depend on the actual measurements while $|T_2|, |T_3|$ are the only quantities that depend on R_1, R_2. Thus, maximizing the likelihood turns into maximizing simultaneously $|T_2|$ and $|T_3|$, or equivalently, maximizing the number of times events (b) and (c) are true. Therefore the components of the optimal R-estimator, $\widehat{R_1}, \widehat{R_2}$ are the solutions of:

$$(\widehat{R_1}, \widehat{R_2}) = argmax_R \parallel X_1 - RX_2 \parallel_0 \tag{13}$$

where the 0-norm means the number of "hits" (i.e. number of cases when $X_1 - RX_2 = 0$). For more than two sources, the additional R's satisfy the same relation, with argmax interpreted as selecting local maxima. The number of local maxima correspond to the number of sources present.

3.3. Implementation of the optimal R-estimator

R-estimates can be obtained by finding the complex R values for every ω and chaining the values together across all frequencies into $R_1(\omega), R_2(\omega), ..., R_n(\omega)$, after appropriate permutations. Below we discuss how: (1) R's are obtained for a particular frequency ω; (2) n is determined from the data at various frequencies; (3) R-estimates are assembled together.

First we address the estimation of R's. Equation 13 can be directly implemented using a histogramming method, as given by the following formula:

$$N_{opt}(R, \omega) = |\{k \ , \ |X_1(\omega, k) - RX_2(\omega, k)| < \delta\}| \tag{14}$$

The optimal estimator is obtained as:

$$\hat{R}_{opt}(\omega) = argmax_R \ N_{opt}(R, \omega)$$

where argmax has the same local maxima interpretation as before.

An approximation of Equation 14 (and therefore suboptimal formula) is obtained further by making explicit the ratio $\frac{X_1(\omega,k)}{X_2(\omega,k)}$:

$$N'(R, \omega) = |\{k \ , \ |\frac{X_1(\omega, k)}{X_2(\omega, k)} - R| < \delta\}| \tag{15}$$

and the R-estimate is constructed similarly to the optimal case, as:

$$\hat{R}'(\omega) = argmax_R \ N'(R, \omega)$$

Note the suboptimal equation above is asymptotically optimal when $\delta \to 0$. Parameter δ corresponds to the bin size used in the computation of histograms $\frac{X_1(\omega,k)}{X_2(\omega,k)}$ for every ω. The determination of local optima naturally corresponds to the highest peaks in the histograms.

Secondly, we return to the question of assessing the number of sources n. Our solution is to select histograms (and, therefore, frequencies) with high confidence peaks. Confidence depends on the height and mass of the peak areas. A reference frequency ω_{ref} is finally obtained, defined as follows: (1) It belongs to a range given by prior knowledge about the type of signals, for instance, 500Hz to 4kHz in the case of voice signals; (2) The minimum distance between acceptable peaks found is the largest among all frequencies considered. The number of peaks for ω_{ref} gives n.

Thirdly, we discuss the way histogram peaks (i.e. R-values) are associated together across all frequencies. For every frequency ω, we consider n histogram peaks and label them $R_{1,2,...,n}(\omega) = \{R_1(\omega), ..., R_n(\omega)\}$. The question is to find a permutation π of the set $1, 2, ..., n$ such that sources for $R_\pi(\omega)$ are in the same order as in $R_{1,2,...,n}(\omega_{ref})$. The optimum permutation $\pi_{opt}(\omega)$ is given by:

$$\pi_{opt}(\omega) = \operatorname{argmax}_\pi \sum_{j=1}^n |\{k \ , \ |\frac{X_1(\omega,k)}{X_2(\omega,k)} - R_{\pi(j)(\omega)}| < \delta \text{ and } |\frac{X_1(\omega_{ref},k)}{X_2(\omega_{ref},k)} - R_j(\omega_{ref})| < \delta\}| \tag{16}$$

Real signals, particularly voice signals can be approximately modeled using our model. Although the discreteness assumption regarding the sources is not valid, for real speech signals the two estimators defined above gave similar results. In conclusion, the suboptimal histogram based estimator is a good estimator for real speech signals, even in the degenerate case of three sources.

3.4. Modeling echoic environments

Imposing a signal mixing model (such as far field and echoic) helps the estimation problem. Here we show how this is done when modeling the environment as an echoic environment of order one (i.e. using only the first indirect path).

We assume that only one source is present, and the distance d between sensors is very small (e.g. microphones are close). Microphone proximity implies that the delays we deal with are fractional. The direct path delays are less than one sample. In the far-field approximation, the transfer functions will have the following form:

$$\begin{aligned} H_{11}(w) &= K(1 + a_1 e^{-i\tau_1 w} + \ldots + a_n e^{-i\tau_n w}) \\ H_{21}(w) &= e^{-i\tau w}(1 + a_1 e^{-i(\tau_1+\delta_1)w} + \ldots + a_n e^{-i(\tau_n+\delta_n)w}) \end{aligned} \tag{17}$$

where $|\tau|, |\delta_1|, \ldots, |\delta_n| < 1$, $a_1, \ldots, a_n < 1$ are the echo attenuations, and $\tau_1, \ldots, \tau_n$ are the echo arrival times.

For $n = 1$ we have an echoic approximation of order one. The model has five parameters in this case: K, τ, τ_1, δ_1 and a_1. Estimation of parameters enables us to demix signals by complex matrix inversion and computation with fractional delays in the two by two problem. It turns out that parameters can be estimated reliably by identification based on SDFT ratios as discussed.

The main steps of the procedure are:

1. Compute SDFT ratios $R(\omega, k)$;
2. Estimate $\hat{R}(\omega, k)$ for each histogram peak;
3. Determine permutation for assembling R-estimates together
4. Identify parameters of the environment model for R's above
5. Recompute R-estimates based on the model
6. Recover independent signals by signal demixing(see next subsection).

Anechoic estimates can be obtained as a particular case of the echoic results, when the echoic model is simplified with $a_1 = 0$. FIR models for H_11 and H_21 can also be used.

3.5. Signal Demixing

In the case of two sources, source estimates are obtained using the adjunct of the inverse of the estimated mixing matrix.

The direct method for signal estimation based on R-estimates in the general case of n sources consists of:

1. Partitioning of the complex plane by a Voronoi tesselation on the set of points $R_{1,2,...,n}$
2. Spectral weighting of mixtures in the frequency domain, with the characteristic function of each Voronoi set
3. Inversion of STFT signals obtained by spectral weigthting.

.

A generalization of the algebraic approach in the degenerate case of more sources than sensors is Wiener filtering, as described below. It requires additional information about the variances of the sources, or a separate estimation process with this goal.

Let us consider the case $n = 3$. Assuming that the variances v_i, $i = 1, 2, 3$, are determined, then an estimate of source i is:

$$\hat{S}_i = \tilde{H}_{1i} X_1 + \tilde{H}_{2i} X_2$$

where:

$$\begin{aligned} \tilde{H}_1 i = \; & (v_1 + v_2 + v_3)(\bar{R}_1 v_1 \delta_{1i} + \bar{R}_2 v_2 \delta_{2i} + \bar{R}_3 v_3 \delta_{3i}) - \\ & -(\bar{R}_1 v_1 + \bar{R}_2 v_2 + \bar{R}_3 v_3)(v_1 \delta_{1i} + v_2 \delta_{2i} + v_3 \delta_{3i}) \\ \tilde{H}_2 i = \; & -(R_1 v_1 + R_2 v_2 + R_3 v_3)(\bar{R}_1 v_1 \delta_{1i} + \bar{R}_2 v_2 \delta_{2i} + \bar{R}_3 v_3 \delta_{3i}) + \\ & (R_1 \bar{R}_1 v_1 + R_2 \bar{R}_2 v_2 + R_3 \bar{R}_3 v_3)(v_1 \delta_{1i} + v_2 \delta_{2i} + v_3 \delta_{3i}) \end{aligned} \tag{18}$$

and δ_{ij} is the Kronecker symbol.

3.6. Generalization of Estimation Approach for More Microphones

The analysis of the case of more sources and two microphones is similar. The ratio histograms will exhibit more local maxima, one peak for each source. Thus the R-estimation problems reduces to finding the peaks of a histogram. The problem becomes more challenging if more sensors are used. Under the deterministic w-disjoint orthogonality no acceptable answer has been found. The obvious solution there would be to pair the microphones and then to somehow average the estimates. Using our stochastic model, a similar computation can be done for the likelihood function. For the three-sensors case, the mixing model is:

$$\begin{aligned} X_1^k &= R_{11} V_1^k G_1^k + R_{12} V_2^k G_2^k + R_{13} V_3^k G_3^k \\ X_2^k &= R_{21} V_1^k G_1^k + R_{22} V_2^k G_2^k + R_{23} V_3^k G_3^k \\ X_3^k &= V_1^k G_1^k + V_2^k G_2^k + V_3^k G_3^k \end{aligned}$$

Assuming the Bernoulli RVs V_1, V_2, V_3 have probabilities $p_1, p_2, p_3 \to 1$ close to one, we obtain:

$$P(X_1, X_2, X_3|R) = const \cdot (1+E_1)^{|T_1|}(1+E_2)^{|T_2|}(1+E_3)^{|T_3|}$$

where

$$T_a = \{k, |X_3^k - \frac{X_1^k(R_{2c}-R_{2b})}{R_{1b}R_{2c}-R_{1c}R_{2b}} - \frac{X_2^k(R_{1b}-R_{1c})}{R_{1b}R_{2c}-R_{1c}R_{2b}}| < \delta\}$$

with (a,b,c) a circular permutation of $(1,2,3)$. Defining $A_a^1 = \frac{R_{2c}-R_{2b}}{R_{1b}R_{2c}-R_{1c}R_{2b}}$ and $A_a^2 = \frac{R_{1b}-R_{1c}}{R_{1b}R_{2c}-R_{1c}R_{2b}}$ and noticing that $\{R_{1a}, R_{2a}; a = 1,2,3\}$ is bijectively mapped onto $\{A_a^1, A_a^2; a = 1,2,3\}$ except for some singularities, it follows we can fist estimate A's parameters and then recover R's. Thus, the optimal estimator is given by the first three optimizers of:

$$(\hat{A_a^1}, \hat{A_a^2}) = argmax_{A_a^1, A_a^2} |\{k, |X_3^k - A_a^1 X_1^k - A_a^2 X_2^k| < \delta\}|$$

If each A is discretized into k bins, the total computationalt cost would be $5Bk^2$. Alternatively, inspired by the ratio estimator in the two-sensor case, a suboptimal estimator can be derived as follows: the sets T_1, T_2, T_3 above are the sets when only one source is zero on the particular frequency ω. The probability of this to happen is about $p^2(1-p)$ much bigger that $p(1-p)^2$ the probability of two sources to be zero. Thus the ratios histogram would have smaller peaks and these peaks would be more difficult to detect. Instead, let us consider the histograms of $\frac{X_1 - z_1 X_3}{X_2 - z_2 X_3}$ for several values z_1, z_2. In general these histograms would exhibit a number of peaks of small amplitude, unless z_1 and z_2 are exactly R_{1a}, R_{2a} for some a. In this case, the contribution of source a is canceled and the ratios histogram would exhibit two big peaks (of amplitude of order $p(1-p)$). These peaks should be at $\frac{R_{1b}-R_{1a}}{R_{2b}-R_{2a}}$, respectively $\frac{R_{1c}-R_{1a}}{R_{2c}-R_{2a}}$, with (a,b,c) a permutation of $(1,2,3)$. This estimator requires the same $5Bk^2$ operations, as the optimal estimator. Further analysis will be done to check the accuracy of this estimator.

4. NUMERICAL RESULTS

We propose a number of tests to show that our model captures well real data.

First we checked our stochastic model (11). For this we took about 18.6 seconds of voice (with natural pauses) at a sampling frequency of $F = 8000Hz$ and computed the STFT coefficients over windows of length $M = 64$. Figure 1 plots the histograms of their real parts when the number of bins is 100. Note the peaks around zero. This is consistent with the superposition formula $Pr = (1-p)P_1 + p * \delta()$ which would represent the p.d.f. of a product VG between a Bernoulli and a continuous random variable.

Next we plot the histograms (14) and (15) of the ratio X_1/X_2 for three sources of type (11) where V's are Bernoulli(p) and G's are uniformly distributed on $[-2,2]$. For $R_1 = -0.5$, $R_2 = 0.5$, $R_3 = -0.1$ and $p = 0.85$, we obtained the histograms drawn in Figure 2. Even though p is close to 1, the peaks can be very well estimated (the smaller the probability p, the higher the peaks).

Finally we examine an echoic environment with one echo as in (16) for two three second TIMIT voices. A source's true complex ratio is drawn with solid line in Figure 3. Using the ratio histogram estimator we obtained the estimates

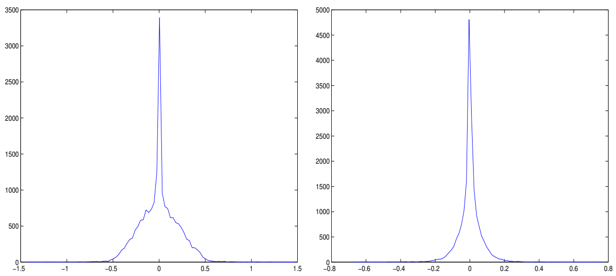

Figure 1: The STFT coefficients histograms for two frequencies ($\omega = 0$ left and $\omega = 15\pi/32$ right).

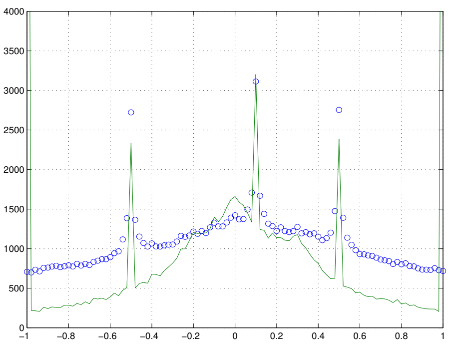

Figure 2: The optimal (dotted line) and DUET (solid line) histograms for a degenerate case (three sources).

drawn with dotted lines in the same figure. For each frequency we estimated the 2 peaks in the histogram, and then we decided how to assign the values based on a continuity property. With the exception of a few frequencies, the estimates are close to the ideal curves. The ratio-estimates obtained from these peaks (i.e. the estimated mixing model parameters) were very close to the true values.

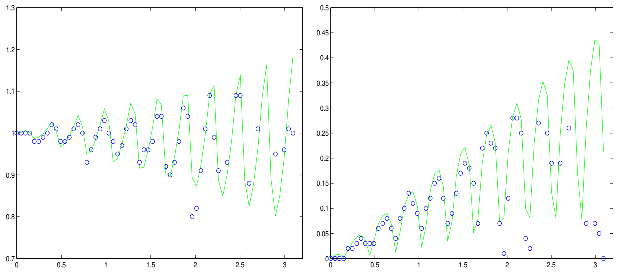

Figure 3: The real (left) and imaginary (right) parts of the estimated ratios (the DUET estimates are the dotted lines).

5. CONCLUSIONS

In this paper we defined and analyzed novel signal processing techniques that can be effectively used for source separation, signal enhancement, and noise reduction when using a twin microphone system. First, we defined the class of stochastic signals for which ratio-estimates can be computed from histograms. This class fits real-world signals of interest such as voice signals. The main theoretical result was the computation, in closed form, of the optimal estimator for this class of signals. Finally we extended the optimal estimator and the DUET suboptimal estimator to the case of more than two channels.

Future work will present variations of the ratio-estimate based algorithms for different environments and problems. Our optimal estimator can be further analyzed to derive identification resolution bounds, and bounds on the number of sources in order to be able to separate sources using ratio-estimate techniques.

Acknowledgment. We thank Joseph Ó Ruanaidh for many useful discussions and proofreading this document.

6. REFERENCES

[1] Radu Balan, Justinian Rosca, Scott Rickard, and Joseph Ó Ruanaidh. The influence of windowing on time delay estimates. In *Proceedings CISS 2000, Princeton, NJ*, 2000. Princeton.

[2] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[3] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE Press, 2000. June 5-9, 2000, Istanbul, Turkey.

[4] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[5] Hamid Krim and Mats Viberg. Two decades of array signal processing research. *IEEE Signal Processing Magazine*, 13(4), 1996.

[6] Tim Laakso, Vesa Valimaki, Matti Karjalainen, and Unto Laine. Splitting the unit delay. *IEEE Signal Processing Magazine*, pages 30–60, 1996.

[7] Lucas Parra, Clay Spence, and Bert De Vries. Convolutive blind source separation based on multiple decorrelation. In *NNSP98*, 1988.

[8] Christian Jutten Pierre Comon and Jeanny Herault. Blind separation of sources, part ii: Problems statement. *Signal Processing*, 24(1):11–20, 1991.

[9] Justinian Rosca, Joseph Ó Ruanaidh, Alexander Jourjine, and Scott Rickard. Broadband direction-of-arrival estimation based on second order statistics. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 775–781. MIT Press, 2000.

[10] K. Torkkola. Blind separation of convolved sources based on information maximization. In *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, 1996.

[11] V. Van Veen and Kevin M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2), 1988.

[12] Ehud Weinstein, Meir Feder, and Alan Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Trans. on Speech and Audio Processing*, 1(4):405–413, 1993.