SCALABLE AUDIO SOURCE SEPARATION IN THE PRESENCE OF NOISE

Justinian Rosca, Radu Balan and Scott Rickard

Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540
{justinian.rosca,radu.balan,scott.rickard}@scr.siemens.com

ABSTRACT

Real blind source separation scenarios are rarely "square" (have equal number of sources as the number of sensors). On the contrary, situations constantly vary between the so called "degenerate" case (more sources than sensors) and the over specified case (less sources than sensors). By being able to deal evenly with such cases and with the presence of noise, the present blind source separation approach opens the door to audio source separation in realistic scenarios. We consider an additive noise mixing model with an arbitrary number of sensors and possibly more sources than sensors, where sources are disjointly orthogonal, i.e. they have mutually disjoint supports of the windowed Fourier transform. The approach estimates the model in the presence of noise under the direct-path far-field assumptions by maximum likelihood. We present the relationship between our solution and classical superdirective beamformers. The implementation of the derived criterion essentially iterates two steps: (1) partitioning of the time-frequency plane for separation, and (2) optimization of the mixing parameter estimates. The solution is applicable to an arbitrary number of microphones and sources, with no prior knowledge about the sources. Experimentally, we showed the capability of the technique to separate four voices from two, four, six and eight channel recordings in the presence of noise.

1. INTRODUCTION

Source separation promises to further a variety of applications of speech enhancement and separation beyond what is possible today with classical microphone array techniques [1]. In particular for audio signals, a variety of BSS techniques have been introduced in recent years. Few work on real audio data (e.g. [2, 3, 4]), even fewer address the noisy case [5], and most deal with the "square" scenario of source separation, i.e. equal number of sources and sensors. Claims of generalization to the non-square case exist, however most often it is not clear how techniques would scale, neither from an algorithmic perspective nor in terms of computational properties.

In this paper we present a novel approach to blind source separation (BSS) exploiting time-frequency (TF) properties of the input data and of the noise, which are readily applied to speech separation on two, four, and six channels. For this, we extend the max-likelihood (ML) estimators derived under the W-disjoint orthogonality assumption in [6]. The ML approach considers both mixing parameters and sources, unlike in [4] where the optimization was over mixing parameters only. We analyze the source and parameter estimators of our model defined below when noise comes from an isotropic diffuse noise field, as studied in differential microphone array literature [1].

2. MIXING MODEL

2.1. Assumptions

Consider the measurements of L source signals by a equispaced linear array of D sensors under far-field assumption where only the direct path is present, where the attenuation and delay parameters of the first mixture $x_1(t)$ are absorbed into the definition of the sources:

$$\begin{aligned} x_1(t) &= \sum_{l=1}^{L} s_l(t) + n_1(t) \\ x_k(t) &= \sum_{l=1}^{L} (1-(d-1)a_l)s_l(t-(d-1)\tau_l) + n_k(t),\ 2 \le k \le D \end{aligned} \quad (1)$$

where $n_1, \ldots, n_D$ are the sensor noises, and $(a_l; \tau_l)$ are average attenuation and delay parameters of source l to sensor array.

We denote by $X_d(k,\omega)$, $S_l(k,\omega)$, $N_d(k,\omega)$ the short-time Fourier transform of signals $x_d(t), s_l(t)$, and $n_d(t)$, respectively, with respect to a window $W(t)$, where k is the frame index, and ω the frequency index. Then the mixing model (1) becomes

$$X(k,\omega) = \sum_{l=1}^{L} Z_l(\omega) S_l(k,\omega) + N(k,\omega) \quad (2)$$

with

$$Z_l(\omega) = \left[\begin{array}{cccc} 1 & (1-a_l)e^{-i\omega\tau_l} & \cdots & (1-(D-1)a_l)e^{-i\omega(D-1)\tau_l} \end{array}\right]^T \quad (3)$$

and X, N the D-vectors of measurements, respectively noises. We assume the noise is Gaussian distributed with a covariance matrix of the form

$$R_n = \sigma^2 \Gamma_n \quad (4)$$

where σ^2 is the average noise field spectral power, and Γ_n the coherence matrix. The uncorrelated noise field is characterized by the identity matrix,

$$\Gamma_n = I_D \quad (5)$$

whereas the isotropic, diffuse noise field has the coherence matrix given in [1].

2.2. BSS Problem

Given measurements $(x_1(t), \ldots, x_D(t))_{1 \le t \le T}$ of the system (1) we want to determine the ML estimates of the mixing parameters $(a_l, \tau_l)_{1 \le l \le L}$ and the source signals $(s_1(t), \ldots, s_L(t))_{1 \le t \le T}$ in the presence of isotropic diffuse noise. When the number of sources is greater than the number of mixtures the problem is degenerate. In order to solve this we rely on a sparseness assumption, W-disjoint orthogonality [7]. L sources are called *W-disjoint orthogonal*, for a given windowing function $W(t)$, if the supports of the windowed Fourier transforms satisfy:

$$S_i(k,\omega) S_j(k,\omega) = 0 \ , \ \forall\, 1 \le i \ne j \le L,\ \forall k, \omega \quad (6)$$

3. THE MAXIMUM LIKELIHOOD ESTIMATOR OF SIGNAL AND MIXING PARAMETERS

Source signals naturally partition the time-frequency plane into L disjoint subsets $\Omega_1, \ldots, \Omega_L$, where each source signal is non-zero (i.e. active). Thus the signals are given by the collection $\Omega_1, \ldots, \Omega_L$ and one complex variable S that defines the active signal:

$$S_l(k,\omega) = S(k,\omega) 1_{\Omega_l}(k,\omega) \tag{7}$$

Let the model parameters θ consist of the mixing parameters (a_l, τ_l), $1 \leq l \leq L$, the partition $(\Omega_l)_{1 \leq l \leq L}$ and S. Based on equations 2 and 4, its likelihood and maximum log-likelihood estimator are given by:

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{l=1}^{L} \prod_{(k,\omega)\in\Omega_l} \frac{1}{\pi^D \sigma^{2D}} \exp\{-\frac{1}{\sigma^2} Y_l^*(k,\omega)\Gamma_n^{-1}(\omega) Y_l(k,\omega)\} \\ \hat{\theta}_{ML} &= \text{argmin}_\theta \sum_{l=1}^{L} \sum_{(k,\omega)\in\Omega_l} Y_l^*(k,\omega)\Gamma_n^{-1}(\omega) Y_l(k,\omega)\} \end{aligned} \tag{8}$$

where $Y_l(k,\omega) = X(k,\omega) - Z_l(\omega) S_l(k,\omega)$. For any partition $(\Omega_1, \ldots, \Omega_L)$ we define the selection map Σ : TF-plane $\rightarrow \{1, \ldots, L\}$, $\Sigma(k,\omega) = l$ iff $(k,\omega) \in \Omega_l$. Clearly Σ defines a unique partition. Optimizing over S in (8) we obtain

$$\hat{S}_{\text{ML}} = \frac{Z_l^* \Gamma_n^{-1} X}{Z_l^* \Gamma_n^{-1} Z_l} \tag{9}$$

where $l = \Sigma(k,\omega)$. Let us denote by $A = (a_l, \tau_l)_{1 \leq l \leq L}$ the mixing parameters. Inserting (9) into (8), the optimization problem reduces to:

$$(\hat{A}, \hat{\Sigma}) = \text{argmax}_{A,\Sigma} J(A, \Sigma) \tag{10}$$

where:

$$J(A,\Sigma) = \sum_{(k,\omega)} \frac{|Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} X(k,\omega)|^2}{Z_{\Sigma(k,\omega)}^* \Gamma_n^{-1} Z_{\Sigma(k,\omega)}} \tag{11}$$

Note the criterion to maximize depends on a set of continuous parameters A, and a selection map Σ. The optimization is done in two steps: first the optimization over the continuous parameters, and then the optimization over the selection map (or, equivalently, the partition). Such a procedure is iterated until the criterion reaches a saturation floor. Because the criterion is bounded above, we are guaranteed it will converge. A description of the solutionj is provided in [8].

4. BSS VS. MICROPHONE ARRAY SOLUTION

The problem of "enhancement" of one source of speech could be also dealt with by means of classical beamforming theory. For instance, one can search for the optimal filter H in $Y = HX$ that maximizes the signal-to-noise-ratio at the output:

$$(\hat{H}) = \text{argmax}_H \frac{E\{|HZ|^2 |S|^2\}}{E\{|HNN^* H^*|\}} \tag{12}$$

It turns out that the solution can be obtained in closed form as follows:

$$\hat{S}_{\text{BF}} = \frac{Z^* \Gamma_n^{-1} X}{Z^* \Gamma_n^{-1} Z} \tag{13}$$

Contrast equation 12 and the solution to the blind source separation problem 9. Apparently there are few differences although the two solutions have been obtained using different assumptions and criteria. However, a closer inpection shows big differences and explains how the BSS problem is actually solved. The ML BSS solution proposes a simultaneous allignment (i.e. beamforming-like) of the signals and TF source separation. We can talk about beamforming being performed in one consistent way only for time-frequency points given by one partition of Σ. At the same time, Σ gives the set of "spatial directions" for multiple source enhancement. Formally, the solution 9 shows that once the mixing parameters have been estimated, we apply two independent linear filters. One linear filter is across the spatial channels (9) and performs a beamforming in order to reduce the output noise. The other (7) is across time-frequency domain and solves the source separation problem by selecting those time-frequency points where, by our W-disjoint orthogonality assumption, only one source is active.

5. CONCLUSIONS

By being able to deal with cases of diffuse noise and more sources than the number of microphones, the present source separation approach opens the door to audio source separation in realistic scenarios. For more details on the technique and a detailed presentation of experimental results with this algorithm the reader is refered to [8]. The resulting algorithm exhibits source separation power and scaling properties both algorithmically (in the number of inputs: we used two, four, six, and eight microphone linear arrays) and experimentally (separation power increases on echoic data with an increase in the number of inputs). In this paper we additionally contrasted the BSS solution obtained with that of a superdireactional microphone array optimizing output signal-to-noise-ratio.

6. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[2] L.Parra, "Convolutive blind source separation based on multiple decorrelation," in *IEEE-ICNN*, 1997.

[3] Jorn Anemuller and Birger Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the second international workshop on independent component analysis and blind signal separation*, Petteri Pajunen and Juha Karhunen, Eds., Helsinki, Finland, June 19–22 2000, pp. 215–220.

[4] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA2001), San Diego, CA*, December 2001, pp. 651–656.

[5] E. Moulines, J.F. Cardoso, and E. Gassiat, "Maximum likelihood for blind source separation and deconvolution of noisy signals using mixture models," in *Proceedings ICASSP*. 1997, pp. 3617–3720, IEEE Press.

[6] R. Balan, J. Rosca, and S. Rickard, "Scalable non-square blind source separation in the presence of noise," in *sent to IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003), Hong-Kong, China*, April 2003.

[7] 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*. 2000, IEEE Press, June 5-9, 2000, Istanbul, Turkey.

[8] R. Balan, J. Rosca, and S. Rickard, "Non-square blind source separation under coherent noise by beamforming and time-frequency masking," in *Submitted to Proc. ICA*, 2003.