IMPROVING MODEL-BASED CONVOLUTIVE BLIND SOURCE SEPARATION TECHNIQUES VIA BOOTSTRAP

Swati Chandna and Wenwu Wang

Centre for Vision, Speech and Signal Processing (CVSSP)
Department of Electronic Engineering (FEPS)
University of Surrey, Guildford GU2 7XH, UK
Email: {s.chandna, w.wang}@surrey.ac.uk

ABSTRACT

Blind source separation for underdetermined reverberant mixtures is often achieved by assuming a statistical model for cues of interest where the unknown parameters of the statistical model depend on hidden variables. Here, the expectation-maximization (EM) algorithm is employed to compute maximum-likelihood estimates of the unknown model parameters. A by-product of the EM algorithm is a time-frequency (T-F) mask which allows the estimation of the target source from the given mixture. In this paper, we propose the idea of bootstrap averaging to improve separation quality from mixtures recorded under reverberant conditions. Our experiments on real speech mixture signals show an increase in the signal-to-distortion ratio (SDR) over a state-of-the-art baseline algorithm, to our knowledge, currently, the best performing technique in this class of methods.

Index Terms— blind source separation, bootstrap averaging, bagging, time-frequency masking, non-stationary multivariate time series, EM algorithm

1. INTRODUCTION

This paper deals with the problem of simultaneously separating multiple sound sources from a two-channel stereo mixture acquired in a reverberant environment. Let $\underline{\mathbf{x}}(n) = [x_1(n), \ldots, x_J(n)]^T$ denote the mixture vector formed by convolutions of the form

$$x_j(n) = \sum_{i=1}^{I} s_i(n) * h_{ij}(n) * e_j(n), \tag{1}$$

where I is the number of sources, J is the number of channels, $s_i(n)$ is the ith source, $h_{ij}(n), j = 1, \ldots, J$ are the room impulse responses from source i to sensor j, and $e_j(n)$ denotes the corresponding noise terms, at discrete time point n. Let $\mathbf{s}_i = [s_i(1), \ldots, s_i(N)]^T$ denote the length-N source vector, and $\mathbf{x}_j = [x_j(1), \ldots, x_j(N)]^T$ denote the corresponding mixture vector. Then, our focus is on the problem of estimating the sources $\mathbf{s}_i, i = 1, \ldots, I$, given mixtures $\mathbf{x}_j$, $j = 1, \ldots, J$. If the number of sources exceed the number of sensors, i.e. $I > J$, the problem is underdetermined, and traditional matrix inversion demixing as in the exact or overdetermined case ($I \leq J$) do not apply. Previously proposed techniques for underdetermined convolutive speech separation have used the assumption that speech signals satisfy the W-disjoint orthogonality (WDO) condition – given speech signals $s_q(n)$ and $s_r(n)$, $q, r \in \{1, \ldots, I\}$, let $S_q(\omega, t)$ and $S_r(\omega, t)$ denote their short-time-Fourier transforms (STFT), respectively, then,

$$S_q(\omega, t) S_r(\omega, t) = 0 \ \ \forall \omega, t \tag{2}$$

i.e. signals have a disjoint support in the time-frequency domain, with ω and t denoting the frequency bin and time frame indices, respectively. The usefulness of this condition in source separation can be understood by considering the simple case where a single-channel mixture ($J = 1$) denoted as $\mathbf{x} = [x(1), \ldots, x(N)]^T$ is formed by a sum of multiple speech signals, for example, with $I = 3$, $\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{s}_3$, which using the STFT leads to $\mathbf{X} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3$, where $\mathbf{X} = [X(\omega, t)] \in \mathbb{C}^{W \times T}$, i.e there are W frequency bins and T time frames, and similarly, $\mathbf{S}_i = [S_i(\omega, t)] \in \mathbb{C}^{W \times T}$. Clearly, now, if the ith source is dominant at (ω, t), for some ω and t, then $X(\omega, t) = S_i(\omega, t)$, and so $\mathbf{s}_i$ can be estimated simply by an inverse STFT of the collection of all such time-frequency points where the ith source dominates.

With this assumption, the problem reduces to identifying the dominant source for each T-F point. This can be done in different ways, for example, in the work done by [1], the ith source is said to be dominant at a given T-F point if its signal-to-interference ratio (SIR) is above a certain threshold; Sawada [2] and Mandel [3], on the other hand perform classification by assuming a statistical model for the mixture vector and interaural cues, respectively, which allows them to compute probabilities for the ith source being dominant at

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307/1 and the MOD University Defence Research Collaboration in Signal Processing.

a given T-F point. Recently, [4] combined the two model-based approaches of [2] and [3], and showed that it leads to a significant increase in SDR. Under reverberant conditions, the WDO assumption is only approximately true, i.e. speech signals begin to partially overlap in the T-F domain as the reverberation time is increased. In this scenario, probabilistic T-F masks as opposed to binary masks, e.g. computed using model-based methods of [2], [3] etc. are known to be useful. However, since no explicit model for reverberation is available, the probability with which a source dominates a given T-F point, computed using the EM algorithm in such methods ([2], [3], and [4]) may not be very reliable. In this paper, we show how separation quality can be improved by bootstrapping the reverberant mixture.

This paper is organized as follows. In section 2, the two well-known model-based expectation-maximization techniques of [2] and [3] are discussed. In section 3 we discuss the bootstrap averaging approach to improve such model-based EM employing source separation techniques. In section 4, we provide results from our experiments on real room-recorded speech mixtures. A brief conclusion is included in section 5.

2. MODEL-BASED EM SEPARATION

Let $\underline{\mathbf{x}}(n) = [x_1(n), x_2(n)]^T$ denote the two-channel mixture vector formed by convolutions of the form (1). For convenience, we shall now use the notation of [3], which deals with the special case of binaural mixtures i.e. $J = 2$. So let $l(n) \equiv x_1(n)$ and $r(n) \equiv x_2(n)$, so that $\underline{\mathbf{x}}(n) = [l(n), r(n)]^T$. Given $\mathbf{l} = [l(1), \ldots, l(N)]^T$, and $\mathbf{r} = [r(1), \ldots, r(N)]^T$, let $\mathbf{L} = [L(\omega, t)] \in \mathbb{C}^{W \times T}$ and $\mathbf{R} = [R(\omega, t)] \in \mathbb{C}^{W \times T}$ denote their STFTs, respectively.

The main idea in [2] is to classify the mixture vector $\underline{\mathbf{x}}(\omega, t) = [L(\omega, t), R(\omega, t)]^T$ for each (ω, t) pair in a frequency bin-wise manner, so that $\underline{\mathbf{x}}(\omega, t)$ is said to belong to a class C_i if the source s_i is most dominant in $\underline{\mathbf{x}}(\omega, t)$. This is done by assuming a complex Gaussian density function for $\underline{\mathbf{x}}(\omega, t)$, i.e., for each ω,

$$\underline{\mathbf{x}}(\omega, t) = \begin{bmatrix} L(\omega, t) \\ R(\omega, t) \end{bmatrix} \equiv \underline{\mathbf{x}}(t) \sim \mathcal{N}^{\mathbb{C}}(\mathbf{a}_i, \gamma_i^2), \tag{3}$$

where $\mathbf{a}_i$ is the mean vector with unit norm and γ_i^2 denotes the variance. Since the total overall log-likelihood of $\sum_{t=1}^{T} \log p(\underline{\mathbf{x}}(t)|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\mathbf{a}_1, \gamma_1, \ldots, \mathbf{a}_I, \gamma_I)$, cannot be maximized directly, the EM algorithm is employed to estimate the unknown parameters in $\boldsymbol{\theta}$. The E-step computes $P(C_i|\underline{\mathbf{x}}(t), \boldsymbol{\theta})$ for each $i = 1, \ldots, I$ and ω. Then, the dominant source is identified based on a comparison of probabilities $P(C_i|\underline{\mathbf{x}}(t), \boldsymbol{\theta})$ at each (ω, t). Since the order of the sources at each frequency bin is not necessarily the same as others, permutation alignment is performed before the signals are transformed to the time domain.

The method proposed in [3] deals with the ratio of $L(\omega, t)$ to $R(\omega, t)$, which is called the interaural spectrogram. This ratio is parameterizable in terms of two interaural cues – the interaural level difference (ILD) denoted by $\alpha(\omega, t)$, and interaural phase difference (IPD) $\phi(\omega, t)$, as follows:

$$\frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega,t)/20} \mathrm{e}^{j\phi(\omega,t)}. \tag{4}$$

A Gaussian distribution is assumed for both $\alpha(\omega, t)$ and $\phi(\omega, t)$, and assuming that points from the same source and at the same delay τ are independently distributed, we are able to write the joint density function of $\alpha(\omega, t)$ and $\phi(\omega, t)$ as

$$\begin{aligned} p(\phi(\omega, t), \alpha(\omega, t)|\Theta) \quad = \quad & p(\phi(\omega, t)|\xi_{i\tau}(\omega), \gamma_{i\tau}^2(\omega)) \times \\ & p(\alpha(\omega, t)|\mu_i(\omega), \eta_i^2(\omega)) \times \\ & p(i, \tau) \end{aligned} \tag{5}$$

where $p(i, \tau) \equiv \psi_{i\tau}$, is the joint probability of any T-F point being in source i at delay τ; and $\Theta = (\xi_{i\tau}(\omega), \gamma_{i\tau}(\omega), \mu_i(\omega), \eta_i^2(\omega), \psi_{i\tau})$ contains the frequency dependent mean and variance parameters of $\phi(\omega, t)$ and $\alpha(\omega, t)$, as well as the joint probability $\psi_{i,\tau}$. Again, the EM algorithm is employed to obtain maximum likelihood estimates (MLEs) of unknown parameters in Θ. Here the E-step computes the conditional probability of the spectrogram point (ω, t) coming from source i and delay τ, given the observed interaural cues $\phi(\omega, t)$ and $\alpha(\omega, t)$ and Θ, i.e., the E-step computes:

$$p((\omega, t) \in \text{source } i, \text{delay } \tau|\phi(\omega, t)\alpha(\omega, t), \Theta) \equiv \nu_{i\tau}(\omega, t), \tag{6}$$

using which MLEs of the unknown parameters are calculated in the M-step ([3, eqn. (18)]). Repeated iterations of the E- and M-steps are performed to obtain final estimates of the parameters and the corresponding $\nu_{i\tau}(\omega, t)$. From the definition of $\nu_{i\tau}(\omega, t)$, it is clear that summing (6) over all possible delays τ is the probability that source i is active at the time-frequency point (ω, t). Therefore, for each source i, a probabilistic T-F mask denoted as $\mathbf{M}_i = [M_i(\omega, t)] \in \mathbb{R}^{W \times T}$ can be obtained as:

$$M_i(\omega, t) = \sum_{\tau} \nu_{i\tau}(\omega, t). \tag{7}$$

This is used to construct estimates of the source signals of interest from the given mixture as mentioned before in section 1. The work done in [4] combines the mixture vector model of [2] with the ILD and IPD models of [3] to estimate the new combined set of parameters denoted as $\Gamma = (\mathbf{a}_i(\omega), \gamma_i(\omega), \xi_{i\tau}(\omega), \gamma_{i\tau}(\omega), \mu_i(\omega), \eta_i^2(\omega), \psi_{i\tau})$. The probabilistic mask obtained as a result of this joint model leads to improvements in separation performance measured by SDR over the two methods of [2] and [3], albeit small, and further improvement in the SDR for reverberant mixtures is desirable. The bootstrap averaging approach to achieve this objective is discussed in the next section.

3. IMPROVING THE T-F MASK

It is clear from eqn. (6) and the discussion following it that the probabilistic mask $\mathbf{M}_i$ given by (7) is derived from the parameter estimates in Θ computed in the final M-step of the EM algorithm. Here Θ corresponds to parameters of the source, whereas, the observed cues contain reverberation effects. This may lead to unreliable parameter estimates, even with the addition of a garbage source as in [3], [4], due to the diffuse character of reverberation. We propose the idea of bootstrap averaging or bagging to replace the maximum likelihood estimates of parameters in Θ computed from one of the EM employing techniques described in section 2, with more robust estimates, and then using the corresponding T-F mask for source separation. The technique of bagging was first suggested in the area of machine learning by Brieman [5]. Although, commonly used in statistical classification and regression problems, it has never been used in the area of blind source separation.

Let $\mathbf{y} = [y_1, \ldots, y_p]^T$ denote a p-variate process with a probability distribution denoted by $F_{\mathbf{y}}(.)$. Let $\mathbf{Y} \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T \in \mathbb{C}^{N \times p}$ denote a sample obtained from $\mathbf{y}$, by independently drawing N samples from the probability distribution $F_{\mathbf{y}}$. Let $\hat{\theta}(\mathbf{Y}) \equiv \hat{\theta}$ denote the statistic of interest derived from $\mathbf{Y}$. Consider the aggregated estimator

$$\hat{\theta}_{\mathrm{A}}(B) = \frac{\hat{\theta}_1 + \ldots + \hat{\theta}_B}{B}, \tag{8}$$

obtained by generating B samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_B$ from $F_{\mathbf{y}}$, and computing $\hat{\theta}_j$ from $\mathbf{Y}_j$. Note that both $\hat{\theta}$ and $\hat{\theta}_{\mathrm{A}}$ are random variables, since different samples from $\mathbf{y}$ will lead to different values of $\hat{\theta}$ and $\hat{\theta}_{\mathrm{A}}$. We show that the mean-squared-error (MSE) of the aggregated estimator $\hat{\theta}_{\mathrm{A}}$ is bounded above by the MSE of $\hat{\theta}$, i.e. $E[(\theta - \hat{\theta}_{\mathrm{A}})^2] \leq E[(\theta - \hat{\theta})^2]$ for a sufficiently large B. Consider

$$\begin{aligned} E[(\theta - \hat{\theta})^2] &= E(\theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2) \\ &= \theta^2 - 2\theta E(\hat{\theta}) + E(\hat{\theta}^2) \\ &\geq \theta^2 - 2\theta E(\hat{\theta}) + \{E(\hat{\theta})\}^2, \end{aligned}$$

since $\mathrm{var}(\hat{\theta}) = E(\hat{\theta}^2) - \{E(\hat{\theta})\}^2 \geq 0$, and which clearly implies that

$$E[(\theta - \hat{\theta})^2] \geq (\theta - E(\hat{\theta}))^2 \approx (\theta - \hat{\theta}_{\mathrm{A}})^2, \tag{9}$$

where a good approximation can be achieved using a sufficiently large B. Then, taking the expected value of the above equation w.r.t. $\mathbf{y}$, we get $E[(\theta - \hat{\theta})^2] \geq E[(\theta - \hat{\theta}_{\mathrm{A}})^2]$. Since in practice, the underlying distribution of the given sample is unknown, we can construct the aggregated estimator by bootstrapping the given sample $\mathbf{Y}$ to obtain $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*$, from which the corresponding bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ can be derived. Then, we define the bootstrap sample version of (8) as

$$\hat{\theta}_{\mathrm{A}}^*(B) = \frac{\hat{\theta}_1^* + \ldots + \hat{\theta}_B^*}{B}. \tag{10}$$

Thus, a smaller mean-squared error estimate can be achieved by using the bootstrap averaged estimator $\hat{\theta}_{\mathrm{A}}^*(B)$ given by (10) provided that we choose a sufficiently large B.

The next step is to incorporate this bootstrap averaged estimate in the source separation framework discussed in section 2. We propose the following: (i) bootstrap the given mixture signal $\underline{\mathbf{X}} = [\mathbf{l}, \mathbf{r}]_{N \times 2}$ to obtain B bootstrap samples $\underline{\mathbf{X}}_1^*, \ldots, \underline{\mathbf{X}}_B^*$, and, (ii) employ one of the model-based EM source separation techniques, e.g. [4] to obtain parameter estimates, $\hat{\Gamma}_1^*, \ldots, \hat{\Gamma}_B^*$. Then the idea is to replace the maximum likelihood parameter estimates in $\hat{\Gamma}$ that were derived from the given mixture signal $\underline{\mathbf{X}}$ by employing the EM algorithm, with the bootstrap averaged estimates i.e. with elements of

$$\hat{\Gamma}_{\mathrm{A}}^* = \frac{\hat{\Gamma}_1^* + \ldots + \hat{\Gamma}_B^*}{B},$$

in the M-step, followed by construction of T-F masks that correspond to this set of averaged estimates. We note that there exists a one-to-one correspondence between the bootstrap model parameter estimates in $\hat{\Gamma}_j^*$ and the bootstrap T-F masks $\mathbf{M}_{i,j}^*, j = 1, \ldots, B$, derived from the B bootstrap mixtures, and therefore, for each source i, the averaged T-F mask defined by $\hat{\mathbf{M}}_{i,\mathrm{A}}^* = [\hat{M}_{i,\mathrm{A}}^*(\omega, t)]$, where

$$\hat{M}_{i,\mathrm{A}}^*(\omega, t) = \frac{\hat{M}_{i,1}^*(\omega, t) + \ldots + \hat{M}_{i,B}^*(\omega, t)}{B}, \tag{11}$$

corresponds to the bootstrap averaged parameter estimates. Now, since the mixture vector $\underline{\mathbf{x}}(n)$ is obtained via convolutions of the form (1), we are dealing with dependent vector time series data in which case simple independent resampling techniques do not apply. We use a recently proposed technique [6] based on circulant embedding to generate statistically similar samples of the given mixture. The simulation methodology of [6] generates samples from nonparametric spectral estimates of the given sample via circulant embedding [7]. The spectral density estimate captures the second-order statistics of the data set and the circulant embedding approach based on circulant matrices makes it a very fast resampling technique.

4. RESULTS

In this section, we present some results with real room-recorded speech mixtures. As in [3] and [4], we use the TIMIT data set from which 15 utterances were selected randomly. From the chosen set of 15 utterances, we randomly consider speech signals $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ to form 15 mixtures using binaural RIRs (BRIRs) measured by Hummersone [8]. These BRIRs are available for different configurations. For example, with the target source placed at 0° and two interfering sources positioned symmetrically on the left and right of the target source at different azimuths denoted by θ°, as well as for a set of five reverberation times RT_{60} corresponding

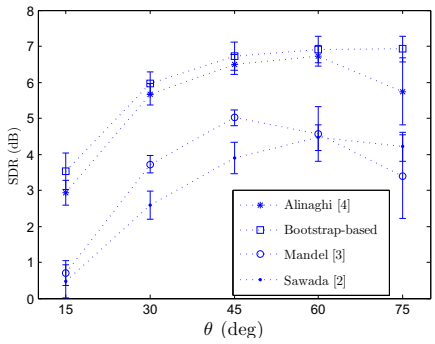

Fig. 1. A comparison of the average SDRs of target sources over a set of 15 two-channel mixtures obtained using the proposed bootstrap improvement on [4], with the joint model of Alinaghi [4], the interaural cue-based technique of Mandel [3], and the mixture model-based method of Sawada [2] for θ = (a) 15°, (b) 30°, (c) 45°, (d) 60°, and (e) 75°, respectively. Error bars show standard error.

to 5 different rooms. We use BRIRs from room D which corresponds to a reverberation time of 0.89s.

One of the key assumptions for the simulation methodology of [6] to work is the stationarity of samples. Since speech signals are highly non-stationary, this technique is not directly applicable. However, since speech signals have a slowly evolving spectrum, it is fairly reasonable to assume that small blocks of length 30ms are stationary. So given $\underline{\mathbf{X}} = [\mathbf{l}, \mathbf{r}]^T$, we consider blocks of $\underline{\mathbf{X}}$ of length 30ms which correspond to $N_b = 480$ samples, where the subscript b indicates the samples in a block. Then following [6], we proceed by estimating multitaper spectral estimates for the bivariate time series data in each block. We employ $K = 8$ sine tapers for multitaper spectral estimation and this leads to a bandwidth of 0.0094. The simulation procedure of [6] is applied to time series in each block which are then put together to get a bootstrap sample of the full-length non-stationary time series $\underline{\mathbf{X}}$. Now, since we are not interested in statistical inference via bootstrap, relatively smaller values for B may be used here. With $B = 500$, we obtain B bootstrap samples $\underline{\mathbf{X}}_1^*, \ldots, \underline{\mathbf{X}}_B^*$, from $\underline{\mathbf{X}}$. Following this, the technique of [4] is employed to obtain parameter estimates and time-frequency masks for each of the B bootstrap samples. So we get $\Gamma_1^*, \ldots, \Gamma_B^*$ and $M_{i,1}^*(\omega, t), \ldots, M_{i,B}^*(\omega, t)$, for each source index i.

The bootstrap averaged T-F mask $\hat{\mathbf{M}}_{i,\mathrm{A}}^*$ is computed as in (11). We perform source separation for the set of 15 two-channel mixtures using this bootstrap averaged T-F mask and compare the average signal-to-distortion ratio (SDR) (over 15 mixtures) with the average SDR obtained from [4], [3] (with a garbage source), and [2], for θ = (a) 15°, (b) 30°, (c) 45°, (d) 60°, and (e) 75° in Fig. 1. The improvement in the SDR of the sources separated using our bootstrap averaged T-F mask and [4] is very clearly visible. The difference between the average SDR from the bootstrap average approach and the approach of [4] is calculated to be (in dB): $0.60, 0.31, 0.23, 0.18, 1.19$ for $\theta = 15°, 30°, 45°, 60°, 75°$, respectively. A t-test confirms that the average gain (in SDR) of 0.5 dB over $15 \times 5 = 75$ mixtures (15 mixtures for each θ) is significant (p-value = 0.0073).

5. CONCLUSION

We have shown how bootstrapping the mixture can improve source separation based on some well-known model-based EM algorithm employing techniques. This paper demonstrates the usefulness of bootstrapping in the area of blind source separation and provides an example to show how it may be used to address other problems in signal processing.

6. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[2] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 139–142.

[3] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 382–394, 2010.

[4] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 209–212.

[5] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[6] S. Chandna and A Walden, "Simulation methodology for inference on physical parameters of complex vector-valued signals," *IEEE Transactions on Signal Processing*, vol. 61, pp. 5260–5269, 2013.

[7] G. Chan and A. T. A. Wood, "Simulation of stationary Gaussian vector fields," *Statistics and Computing*, vol. 9, pp. 265–268, 1999.

[8] C. Hummersone, *A psychoacoustic engineering approach to machine sound source separation in reverberant environments*, Ph.D. thesis, University of Surrey, UK, 2011.