AN EXPERIMENTAL COMPARISON OF SOURCE SEPARATION AND BEAMFORMING TECHNIQUES FOR MICROPHONE ARRAY SIGNAL ENHANCEMENT

Joachim Thiemann ∗
Carl-von-Ossietzky University Oldenburg
D-26129 Oldenburg, Germany
joachim.thiemann@uni-oldenburg.de

Emmanuel Vincent
Inria, Centre de Nancy - Grand Est
F-54600 Villers-lès-Nancy, France
emmanuel.vincent@inria.fr

ABSTRACT
We consider the problem of separating one or more speech signals from a noisy background. Although blind source separation (BSS) and beamforming techniques have both been exploited in this context, the former have typically been applied to small microphone arrays and the latter to larger arrays. In this paper, we provide an experimental comparison of some established beamforming and post-filtering techniques on the one hand and modern BSS techniques involving advanced spectral models on the other hand. We analyze the results as a function of the number of microphones, the number of speakers and the input Signal-to-Noise Ratio (iSNR) w.r.t. multichannel real-world environmental noise recordings. The results of the comparison show that, provided that a suitable post-filter or spectral model is chosen, beamforming performs similarly to BSS on average in the single-speaker case while in the two-speaker case BSS exceeds beamformer performance. Crucially, this claim holds independently of the number of microphones.

Index Terms: Source separation, beamforming, FASST, MVDR, post-filtering, evaluation

1. INTRODUCTION
In the signal processing community, blind source separation (BSS) [1–4] and beamforming [5, 6] are typically regarded as distinct families of techniques. Although they share the goal of isolating a target signal from a mixed signal, beamforming relies on the spatial location of the target source, while BSS relies on more general statistical properties of the source signals.

With few exceptions, e.g., [7, 8], BSS has typically been applied to noiseless single-channel or two-channel recordings in the past. In contrast, beamforming has often been employed with larger microphone arrays in noisy conditions, with each increase in the number of channels improving the ability to enhance the target. Recently, however, the frontier between these application domains has blurred and both families of techniques have been tried on mixed setups, e.g., two-channel mixtures of speech and real-world noise [9–11].

This raises the question of the relative performance of BSS and beamforming as a function of the number of microphones, the amount of noise, and other acoustic conditions. A few comparisons have been made that focus on the separation of determined noiseless mixtures of speech via early Independent Component Analysis (ICA)-based techniques [12, 13] or separation of recordings with a slightly larger but fixed number of channels [14]. Yet, to the best of our knowledge, no experiments have been performed to compare modern beamforming techniques involving post-filtering [15–17] with state-of-the-art BSS techniques including advanced spectral modelling [18–20] for different numbers of channels and acoustic conditions.

This paper aims to provide such an experimental comparison for the extraction of one or two target sources mixed with real-world environmental noise recorded by 2 to 8 microphones. We consider the case where the spatial location of the target source(s) is precisely known. For beamforming, this is essential for steering the beam in the correct direction. In BSS, this is used to initialize the estimation process so that separation can be performed without prior knowledge about the spectra of the speech sources. However, we show that the addition of constraints over these spectra improves performance over the use of spatial information alone. Such a setup is sometimes called Semi-Blind Source Separation (SBSS) [10].

In Section 2, we describe the signal models and the estimation algorithms underlying the chosen beamforming and BSS algorithms. In Section 3, we present the experimental setup and the resulting source separation performance. We conclude in Section 4.
is computed that accounts for the Time Differences Of Arrival (TDOA) of a given source \( j \) between the microphones of the array, where \( \tau_{mn}, m = 1, \ldots, M \) is the time it takes for the sound to travel from source \( j \) to microphone \( m \) on a direct path. A set of weights \( w_{j,fn}^H \) are then computed from \( x_{fn} \) and \( d_{j,f} \) and applied to the mixture such that

\[ \hat{s}_{j,fn} = w_{j,fn}^H x_{fn}. \] (3)

In the following, we consider the Minimum Variance Distortionless Response (MVDR) beamformer [21], which minimizes the energy of the interfering sources and noise under the constraint of unitless Response (MVDR) beamformer [21], which minimizes the energy of the interfering sources and noise under the constraint of unitless Response (MVDR) beamformer [21], which minimizes the energy of the interfering sources and noise under the constraint of unitless Response (MVDR) beamformer [21], which minimizes the energy of the interfering sources and noise under the constraint of unit

where the gain factor \( \alpha \) prevents instabilities [22].

\[ w_{j,fn} = \frac{(R_{\alpha,fn} + \alpha I)^{-1} d_{j,f}}{d_{j,f}^H (R_{\alpha,fn} + \alpha I)^{-1} d_{j,f}}, \] (4)

where \( I \) is the \( M \times M \) identity matrix and the diagonal loading factor \( \alpha \) prevents instabilities [22].

2.1.2. Postfiltering for beamformers

Since beamforming is only capable of spatial filtering, an additional single-channel post-filter is often applied which modifies (3) to

\[ \hat{s}_{j,fn}^{\text{post}} = p_{n,f} w_{j,fn}^H x_{fn}, \] (5)

where the gain factor \( p_{n,f} \) is real-valued and typically limited to \( 0 \leq p_{n,f} \leq 1 \) [23]. Such post-filtering provides a fair comparison of beamforming and source separation algorithms which also utilize some form of spectral filtering. The combination of MVDR beamforming with the ideal post-filter equal to the ratio of the (unknown) short-term power spectrum of the target and that of the beamformer output has been shown to estimate the target source in the Minimum Mean Square Error (MMSE) sense [15].

To provide a contrasted comparison, we evaluate three very diverse choices:

- no post-filtering (MVDR no post)
- Zelinski’s post-filter [16], which approximates the ideal post-filter by estimating the noise power spectrum from the off-diagonal values of the observed cross-channel covariances under a diffuse noise assumption (MVDR Zelinski),

Alternative post-filters have since been proposed [15, 17, 23] which we evaluated in preliminary experiments. McCowan’s post-filter performed more poorly than Zelinski’s in the considered acoustic setting. Ito’s post-filter achieved a significant gain over Zelinski’s, but it is less established and did not reach the performance of OM-LSA/IMCRA. Due to space restrictions, we only provide the results of OM-LSA/IMCRA and Zelinski’s post-filter in the following experiments.

2.2. Variance model-based source separation

In contrast to beamforming, BSS estimates all sources jointly. A number of techniques have been proposed that are based either on spatial or on spectral models of the sources [2]. We select here the state-of-the-art flexible variance model in [20] which, together with [26], is one of the few models able to jointly exploit spatial and spectral cues. Its advantage compared to [26] is that it can impose constraints on the spectra of the sources while learning them from the mixture signal.

This method relies on the local Gaussian model

\[ y_{j,fn} \sim N(0, v_{j,fn} R_{j,f}), \] (6)

where \( v_{j,fn} \) represents the short-term power spectrum of the \( j \)-th source and \( R_{j,f} \) its spatial covariance matrix. In order to provide a contrasted comparison again, we consider two different models for the power spectra of the speech sources:

- either \( v_{j,fn} \) is unconstrained (BSS unconstr.),
- or it is modelled via harmonic Nonnegative Matrix Factorization (NMF) [27] (BSS harmonic)

\[ v_{j,fn} = \sum_{k=1}^{L} \sum_{l=1}^{K} w_{j,fl} u_{j,lk} P_{j,kn} \] (7)

where \( w_{j,fl} \) are fixed narrowband spectral patterns with either voiced or unvoiced fine structure and \( u_{j,lk} \) and \( P_{j,kn} \) are spectral envelope and time activation coefficients to be estimated from the observed signal.

Analogous to the steering vector for the beamforming methods, the spatial covariance matrices of the speech sources are computed from \( d_{j,f} \), as described in more detail in the following section. For the noise source, the spatial covariance matrix is initialized with the identity matrix. All spectral parameters are randomly initialized. The noise spatial covariance matrix and all spectral parameters are estimated in the Maximum Likelihood (ML) sense via an iterative Expectation-Maximization (EM) algorithm. The final estimates of the sources are computed in the MMSE sense via the Wiener filter

\[ \hat{y}_{j,fn} = v_{j,fn} R_{j,f} \left( \sum_{j=1}^{I} v_{j,fn} R_{j,f} \right)^{-1} x_{fn}. \] (8)

3. COMPARISON FRAMEWORK AND RESULTS

3.1. Acoustic setup

Our experimental setup is intended to simulate a somewhat realistic usage scenario. One or two target speech signals are convolved by room impulse responses simulated via the source image technique and added to field recordings of environmental noise. The microphone array setup, the room dimensions and the reverberation time in the simulated environment mirrors the physical array with which the background noise was recorded. The use of simulated room impulse responses is widespread in source separation evaluations and has been shown to result in similar separation performance as actual microphone recordings [14].

For the purpose of this evaluation, we collected a database of 16-channel noise recordings (DEMAND) [28] from 18 real-world environments. This study uses a subset of the original recorded channels corresponding to a planar array of 8 microphones in three staggered rows such that the distance from each microphone to its immediate neighbour is 5 cm. The simulated target signal sources are placed one meter away from microphone one of the array, separated by 45 degrees, in the same plane as the array, 1.5 m off the ground (Fig. 1). To gauge how the algorithms scale with the number of available channels, proper subsets of the array are also evaluated.
Regarding the BSS algorithms, we use the FASST toolbox\(^2\) up to some modifications required to account for the increased number of channels compared to the reference two-channel implementation. In all experiments, the noise is modelled as a single source with a full-rank adaptive spatial covariance matrix and spectral parameters constrained to NMF with 16 components. For the target sources, the spatial covariance matrices are fixed to \(R_{x,f} = d_{j,f} d_{j,f}^H + \sigma^2 \Omega_f\), where \(\Omega_f\) is the theoretical covariance matrix of a diffuse noise, as detailed in [29]. The reverberant-to-direct ratio \(\sigma^2\) is set to a small value (0.001) since the algorithm is not given any prior information about the reverberant condition of the signal. For the BSS harmonic experiments, the harmonic constraint (7) is implemented by fixing the narrowband spectra \(w_{j,f}\) as in [20]. To ensure the EM algorithm has fully converged, 100 iterations per channel are used resulting in computational complexity several magnitudes higher than the beamforming methods. Typically a much lower number of iterations can be used with little loss of performance, but a detailed analysis of complexity is outside the scope of the current study.

### 3.3. Evaluation of separation performance

Rejection of interfering sources and noise was evaluated in terms of the Signal-to-Interference Ratio (SIR) as computed by BSS Eval [30], where the background noise and the second source (if present) are both considered as interfering sources. We compare the various algorithms in terms of the SIR improvement (SIRI) with respect to an appropriate “null” algorithm, a metric similar to the noise reduction rate in [13]. In the beamforming case this “null” algorithm simply outputs the signal from the first microphone (located at the left centre of the array). In the source separation case, the “null” algorithm produces the input signals unmodified. Both “null” algorithms provide very similar SIR. This provides a consistent baseline for all algorithms in both the one and two target source cases.

To evaluate signal distortion, the Mel-cepstral distance (MCD) [31] measure is used, defined as

\[
\text{MCD}_{j,n} = \sqrt{\sum_{i=1}^{12} (\text{MC}_{j,i,n} - \text{MC}_{s,j,i,n})^2},
\]

where \(\text{MC}_{s,j,i,n}\) and \(\text{MC}_{j,i,n}\) represent the \(i\)th Mel-frequency cepstral coefficient of the estimated and clean speech source \(j\) in time frame \(n\) respectively, computed using [32]. In the case of BSS, only the first channel of the estimates \(\hat{y}_{j,fn}\) is used to compute the MCD. The per-signal value is averaged over all time frames and all sources. The overall scores shown in the figures below are obtained by averaging SIRI and MCD respectively over all 10 noise environments.

### 3.4. SIRI for a single target in a noise field

Fig. 2a shows the SIRI resulting from the various algorithms in the single target scenario for a 4-channel array at different iSNRs. We find in this scenario that the best performance can be achieved with either MVDR OMLSA or BSS harmonic. BSS unconstr. performs favourably compared with MVDR and MVDR Zelinski, but cannot approach the methods exploiting spectral cues.

In Fig. 2b we show how the SIRI varies with respect to the number of channels, for an iSNR of 0dB. We find that the BSS algo-

---

\(^1\)http://webeer.technion.ac.il/Sites/People/IsraelCohen/Download/omlsa.m

\(^2\)http://bass-db.gforge.inria.fr/fasst/

\(^3\)This metric is also sometimes termed Signal-to-Interference-plus-Noise Ratio (SINR).

---

**Table 1.** Summary of noise recordings used in this study, the characteristics of the noises and the reverberation time of the corresponding room simulation. “Periodic” indicates repeated changes (e.g., typing noises), “mixed” indicates stationary noise interspersed with brief different noises. For the room simulation characteristics, “open” areas are characterized by having direct path reflections only, without significant diffuse reverberation.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Principal noise character</th>
<th>Room sim.</th>
<th>(T_{60})</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFIELD</td>
<td>stationary</td>
<td>open</td>
<td></td>
</tr>
<tr>
<td>NFIELD</td>
<td>periodic</td>
<td>open</td>
<td></td>
</tr>
<tr>
<td>OHALLWAY</td>
<td>stationary</td>
<td>open</td>
<td>0.16 s</td>
</tr>
<tr>
<td>OOFFICE</td>
<td>periodic</td>
<td>open</td>
<td>0.16 s</td>
</tr>
<tr>
<td>PRESTO</td>
<td>babble</td>
<td>open</td>
<td>0.76 s</td>
</tr>
<tr>
<td>PSTATION</td>
<td>stationary</td>
<td>open</td>
<td>0.76 s</td>
</tr>
<tr>
<td>SPSQUARE</td>
<td>mixed</td>
<td>open</td>
<td></td>
</tr>
<tr>
<td>STRAFFIC</td>
<td>stationary</td>
<td>open</td>
<td>0.04 s</td>
</tr>
<tr>
<td>TCAR</td>
<td>stationary</td>
<td>open</td>
<td>0.05 s</td>
</tr>
<tr>
<td>TEMETRO</td>
<td>mixed</td>
<td>open</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Relative position of the simulated sources to the microphone array and geometry of the microphones within the array (not to scale). The microphone array and sources are located 1.5 m off the ground.
3.5. SIRI for two simultaneous targets in a noise field

In the two-target scenario, the BSS algorithms perform significantly better than beamforming, as shown in Figs. 2c and 2d. Again, additional channels confer a great benefit to BSS. Since the beamforming algorithms only consider one direction (of one target), the other target is considered part of the background noise and cannot be suppressed as effectively. The performance of MVDR OMLSA is drastically reduced when compared to the single target case, due to the non-stationary nature of the interference. Overall, the best results are achieved by BSS harmonic for all iSNRs and all numbers of channels.

3.6. Signal distortion

The MCD for the resulting speech signals is shown in Figs. 3a and 3b for the two-channel case. With the exception of MVDR Zelinski, the results do not vary significantly as the number of channels is increased. We find that the distortion incurred by BSS unconst. and the plain MVDR beamformer is about equivalent for both the one- and two-target case. BSS harmonic results in larger distortion especially at high iSNRs since speech sources fit the imposed voiced/unvoiced spectral constraints only to a certain extent. MVDR OMLSA adds further signal distortion due to the aggressive removal of the noise that approaches a binary mask, but can also add musical noise like artifacts. Finally, MVDR Zelinski performed surprisingly poorly, which we attribute to the strong low-pass filtering effect that is observed in the resulting signal for most of the environments. When using 4 or 8 channels, the low-pass effect was even stronger, increasing the MCD further.

4. CONCLUSION

We provide a comparison of some recent source separation and beamforming techniques in a scenario involving either one or two target speech sources in a noise field. We analyze the results as a function of the number of channels, the number of speakers and the iSNR. We find that beamforming with a very effective post-filter performs similar to source separation with a harmonic spectral model only in the single target case. In the case where there are two speech signals interfering with each other, source separation has a clear advantage. However, we note that the iterative nature of the EM algorithm incurs a high computational cost, limiting the scenarios in which these algorithms can be applied. Future work will consider a wider range of techniques and analyze complexity and the impact of additional algorithm parameters.

5. REFERENCES
