Distributed speech separation in spatially unconstrained microphone arrays
Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid

To cite this version:
Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid. Distributed speech separation in spatially unconstrained microphone arrays. ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing, Jun 2021, Toronto / Virtual, Canada. 10.1109/ICASSP39728.2021.9414758. hal-02985794v3

HAL Id: hal-02985794
https://hal.science/hal-02985794v3
Submitted on 15 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
DISTRIBUTED SPEECH SEPARATION IN SPATIALLY UNCONSTRAINED MICROPHONE ARRAYS

Nicolas Furnon, Romain Serizel, Irina Illina

Université de Lorraine, CNRS, Inria, Loria
F-54000 Nancy, France
{firstname.lastname}@loria.fr

Slim Essid

LTIC, Télécom Paris,
Institut Polytechnique de Paris,
Palaiseau, France
slim.essid@telecom-paristech.fr

ABSTRACT

Speech separation with several speakers is a challenging task because of the non-stationarity of the speech and the strong signal similarity between interferent sources. Current state-of-the-art solutions can separate well the different sources using sophisticated deep neural networks which are very tedious to train. When several microphones are available, spatial information can be exploited to design much simpler algorithms to discriminate speakers. We propose a distributed algorithm that can process spatial information in a spatially unconstrained microphone array. The algorithm relies on a convolutional recurrent neural network that can exploit the signal diversity from the distributed nodes. In a typical case of a meeting room, this algorithm can capture an estimate of each source in a first step and propagate it over the microphone array in order to increase the separation performance in a second step. We show that this approach performs even better when the number of sources and nodes increases. We also study the influence of a mismatch in the number of sources between the training and testing conditions.

Index Terms— Speech separation, microphone arrays, distributed processing.

1. INTRODUCTION

Speech separation aims at extracting the speech signals of each speaker in a noisy mixture. It has many applications, for example in automatic speech recognition [1], hearing aids [2] or music processing [3]. In recent years, deep neural network (DNN)-based solutions have replaced model-based approaches because of the great progress they enabled [4–8]. However, most of these DNN-based solutions are developed in “clean” contexts, where the speech signals are not corrupted by noise or by reverberation, which makes them quite unrealistic for real-life applications. Besides, the recent trend shows that state-of-the-art results are mostly achieved with complex DNNs [6–9], with very high model sizes [10]. They also operate on single-channel data, and hence neglect the spatial information which could be accessible in everyday life scenarios, where the majority of recording devices are embedded with multiple microphones. Some solutions have been designed to address the case of multichannel scenarios [11,12] or reverberant conditions [13], but they result in even more complex DNNs than those used in the single-channel context. In everyday life scenarios, where several people speak in presence of their personal devices, these solutions are not (yet) applicable because the computational requirements are too demanding for the memory, computing and energy capacities of the devices.

One way to reduce the computational cost of the DNN-based methodologies while exploiting spatial information is to use ad-hoc microphone arrays and to distribute the processing over all the devices of the array. In a previous article, we introduced a solution that proved to efficiently process multichannel data in a distributed microphone array in the context of speech enhancement [14]. This approach was based on a two-step version of the distributed adaptive node-specific signal estimation (DANSE) algorithm by Bertrand and Moonen, where so-called compressed signals are sent among the devices [15]. It achieved, with a simple convolutional recurrent neural network (CRNN), competitive results compared to a filter computed from oracle time-frequency (TF) masks. In an extended study, we have shown that exchanging the estimates of both the target and the noise across the devices could potentially increase the final performance [16].

In this paper, we show that this solution can be adapted to speech separation in a typical case of a meeting, where all the sources must be estimated in a reverberant but noise-free environment. We redesign the solution such that each device sends an estimate of a different source in a first step. This way, in a second step, each device has access to proper insight into the whole acoustic scene and can separate the signals more accurately. We analyse the performance over the number of sources present and devices available.

This paper is organised as follows. We formulate the problem in Section 2 and detail our solution in Section 3. The experimental setup on which we test the solution is described in Section 4. The results are given in Section 5. We conclude the paper in Section 6.

2. PROBLEM FORMULATION

We consider a scenario where \( N \) speakers \( \{s_n\}_{n=1..N} \) are recorded by \( K \) devices, thereafter called nodes. Each node \( k \) contains \( M_k \) microphones. We assume that no noise is present, so the \( m\)-th micro-

---

This work was made with the support of the French National Research Agency, in the framework of the project DiSCogs “Distant speech communication with heterogeneous unconstrained microphone arrays” (ANR-17-CE23-0026-01). Experiments presented in this paper were partially carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000).
phone of the $k$-th node records the following signal at a time stamp $t$:

$$y_{mk}(t) = \sum_{n=1}^{N} c_{n,m,k}(t) * s_{n}(t)$$

$$= \sum_{n=1}^{N} \hat{s}_{n,m,k}(t)$$

where $*$ stands for the convolution operator and $c_{n,m,k}$ is the room impulse response (RIR) from the source $n$ to the microphone $m_k$ such that $\hat{s}_{n,m,k} = c_{n,m,k} * s_{n}$ is the reverberated image of the source $n$ captured by the $m_k$-th microphone. Speech separation aims at recovering all the speech signals $\{\hat{s}_{n,m,k}\}_{n=1...N}$ where $\mu^{(n)}$ is the reference microphone for source $s_n$.

In the short-time Fourier transform (STFT) domain, we can write:

$$y_{mk}(f, \tau) = \sum_{n=1}^{N} \hat{y}_{n,m,k}(f, \tau)$$

where $f$ is the frequency index and $\tau$ the frame index, and where $y$ and $\hat{y}$ are the STFTs of $y$ and $\hat{y}$ respectively. In the sequel, for the sake of conciseness, we will omit the time and frequency indexes. The under bar indicates a signal in the STFT domain.

In the context of ad-hoc microphone arrays, we can gather all the speech signals of all nodes into another vector:

$$y = \left[ y_{11}^T, \ldots, y_{K1}^T, \ldots, y_{1N}^T, \ldots, y_{KN}^T \right]^T.$$ 

In the second step, a second MWF $w_{kk}$ is applied on $y_k$ to finally estimate the source $\hat{s}_{k,\mu(k)}$:

$$\hat{s}_{\mu k} = w_k^H \hat{y}_k.$$ 

4. SETUP

4.1 Dataset

We simulate a typical situation of a meeting, where $N$ persons are talking around a round table, each of the speakers having laid their smartphone, tablet or computer in front of them on the table. We assume that all the nodes have $M_k = 4$ microphones. The shoebox-like room has a random length, width and height uniformly drawn in $[3, 9]$ meters, $[3, 7]$ meters and $[2.5, 3]$ meters, respectively. The table has a radius randomly drawn between 0.3 m and 2.5 m. The height of the table is randomly selected between 0.8 m and 0.9 m.
Step 1 convolutional layers have 32, 64 and 64 filters respectively, with kernel followed by a recurrent layer and a fully-connected layer. The computation of the IRM at node $k$ is as follows:

$$\text{IRM}_k = \frac{|\hat{s}_{k,\mu}(k)|}{|\hat{s}_{k,\mu}(k)| + |\hat{s}_{k,\mu}(k)|}$$

where $\mu$ is the reference microphone for source $s_k$ and $|\hat{s}_{k,\mu}(k)|$ is the signal statistics at node $k$.

This oracle method is denoted “IRM” in the legends of the following figures. The second method, denoted “MN”, is our multi-node solution where the CRNN sees at the second step the compressed signals to predict the masks. The third method, denoted “SN”, is the single-node solution where the same CRNN sees only the local signal to predict the masks at both filtering steps. The fourth method, denoted “MWF”, is a MWF applied on each node without exchanging the signals. We analyse the behaviour of these methods when the number of sources and nodes varies.

### 5.1. Performance with an equal number of sources

In this section, the four methods are compared in the scenarios where the number of sources and nodes is equal. The performance in terms of SI-SDR is reported in Figure 4. First, although the single-node solution and the MWF do not differ much, the single-node distributed processing shows a significant improvement over the MWF when the number of nodes increases. This shows that exploiting the spatial information conveyed by the whole microphone array helps improving the separation performance. Besides, the multi-node solution significantly outperforms both the single-node solution and the MWF. This shows that the compressed signals are useful not only for the beamforming but also for the mask prediction. Finally, the SI-SDR increases when the number of nodes and sources increases, even if the task gets more challenging, and achieves less than 0.5 dB worse than the oracle performance. This is because the output performance remains constant while the input SI-SDR decreases. This shows the robustness of our solution to spatial diversity.

### 5.2. Performance in over-determined cases

In this section, we analyse the performance of the previous methods in the case where the number of nodes is greater than the number of sources. This could happen in a real situation, for example if a person gets out of the room while leaving their phone on the table. In such a scenario, each node estimates the source in front of which it is placed, which means that the node left without speaker has no target speaker. At this node, the target and noise components of
Equation (4) are not defined, and it is not obvious to determine the compressed signal to send with the IRM method. Because of this, we omit the results obtained with the IRM in this section. The results of the three other methods are reported in Figure 5 where we recall the performance of the equally-determined case ($N = K = 2$ and $N = K = 3$) for an easier comparison.

In over-determined cases as well, the multi-node solution outperforms almost always the other two methods. An interesting exception can be noticed when there are two more nodes than sources ($N = 2$, $K = 4$). We observed that the nodes placed in front of no source estimate completely silent compressed signals, because the masks predicted by the single-node CRNN are close to 0 in almost all TF bins. This means that the filter at the second step is applied on silent signals, which degrades the final performance. Since the multi-node solution still outperforms the single-node solution, the silent compressed signals do not degrade the mask prediction. However, they degrade the distributed beamforming output. This effect is not dominant when there is only one more node than the number of sources, which means that our solution is robust to a source mismatch between training and testing, but only to a limited extent. One solution to cope with this performance drop, could be to automatically ignore the recording devices which start to send silent signals.

Finally, the performance consistently increases for all methods when the number of nodes increases from $K = 2$ to $K = 3$ with two sources. This shows that the neural networks trained with more sources (so on harder conditions) perform better than those trained with a lesser number of sources.

5.3. Performance in under-determined cases

We now focus on the performance of the proposed methods in the case where the number of nodes is smaller than the number of sources. This could happen in a real situation, for example if a phone shuts down during the meeting. The performance for the scenario with $N = 3$ sources is presented in Figure 6 where we recall the performance of the equally-determined case ($N = K = 3$) for an easier comparison.

In this under-determined case, even if there is only one more source than nodes, the multi-node solution performs worse than the two other ones. The mismatch between training and testing leads to worse performance. Since the single-node solution, as well as the MWF, performs quite well, it means that the drop of performance of the multi-node solution is due to the multi-node CRNN that is trained on mixtures with only one interferent speaker while tested on mixtures with two interferent speakers. A similar behaviour was observed in the under-determined cases with $N = 4$ sources.

This indicates that dealing with under-determined cases probably requires to train specific networks with the proper number of interferent sources. Training and testing on a variable number of sources remains an open challenge. This could be addressed within our spatially distributed framework by adapting a dedicated strategy, e.g. in estimating the sources iteratively [21] or by adapting the loss function [22, 23].

6. CONCLUSION

We introduced a distributed algorithm that can process spatial information captured by a spatially unconstrained microphone array. The spatial information is propagated over the microphone array to deliver to all the nodes a global insight into the whole acoustic scene. We evaluated the algorithm in typical meeting configurations and showed that thanks to the spatial information, a CRNN can predict accurate TF masks which lead to almost oracle performance. In scenarios where the number of nodes match the number of sources, we showed that the performance increases when the number of sources (and nodes) increases. We also analysed the limits of this approach when the number of nodes does not match the number of sources. Solving this problem could require the adaptation of techniques developed in the case of fixed microphone arrays.
7. REFERENCES


