



# DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays

Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid

## ► To cite this version:

Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid. DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays. 2019. hal-02389159v1

**HAL Id: hal-02389159**

**<https://hal.science/hal-02389159v1>**

Preprint submitted on 2 Dec 2019 (v1), last revised 11 Mar 2020 (v3)

**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.

# DNN-BASED DISTRIBUTED MULTICHANNEL MASK ESTIMATION FOR SPEECH ENHANCEMENT IN MICROPHONE ARRAYS

Nicolas Furnon, Romain Serizel, Irina Illina

Slim Essid

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

Télécom ParisTech, Univ. Paris-Saclay, LTCI  
75013 Paris, France  
slim.essid@telecom-paristech.fr

## ABSTRACT

Multichannel processing is widely used for speech enhancement but several limitations appear when trying to deploy these solutions to the real-world. Distributed sensor arrays that consider several devices with a few microphones is a viable alternative that allows for exploiting the multiple devices equipped with microphones that we are using in our everyday life. In this context, we propose to extend the distributed adaptive node-specific signal estimation approach to a neural networks framework. At each node, a local filtering is performed to send one signal to the other nodes where a mask is estimated by a neural network in order to compute a global multichannel Wiener filter. In an array of two nodes, we show that this additional signal can be efficiently taken into account to predict the masks and leads to better speech enhancement performances than when the mask estimation relies only on the local signals.

**Index Terms**— Speech enhancement, microphone arrays, distributed processing.

## 1. INTRODUCTION

Almost all voice-based applications such as mobile communication, hearing aids or human to machine interfaces require a clean version of speech for an optimal use. Single-channel speech enhancement can substantially improve the speech intelligibility and speech recognition of a noisy mixture [1, 2]. However improvement with a single-channel filter is limited by the distortions introduced during the filtering operation. The distortion can be reduced in multichannel processing, relying on several microphones, that account for exploiting spatial information [3, 4]. The multichannel Wiener filter (MWF) [5] for example yields the optimal filter in the mean squared error (MSE) sense and can be extended to a speech distortion weighted multichannel Wiener filter (SDW-MWF) where the noise reduction is balanced by the speech distortion [6].

Up to a certain point, the effectiveness of these algorithms increases with the number of microphones. More microphones can allow for a wider coverage of the acoustic scene and a more accurate measure of the statistics of the source signals. In large rooms, or even in flats, this implies the need of huge microphone arrays, which, if they are constrained, can become prohibitive, expensive and lacks flexibility. However, in our daily life, we are surrounded by

an increased number of embedded microphones with the omnipresence of computers, telephones and tablets. They can be viewed as unconstrained, ad hoc wireless acoustic sensor networks (WASNs) with promising potential but also challenges [7]. A distributed adaptive node-specific signal estimation (DANSE) algorithm [8] was proposed for a fully connected network, converging to the centralized MWF [9], where the nodes exchange a single linear combination of their local signals. The constraint of a fully-connected network can be lifted with randomized gossiping-based algorithms, where beamformer coefficients are then computed in a distributed fashion [10]. Message passing [11] or diffusion-based [12] algorithms can increase the rather slow convergence rate of these solutions. Another way to exploit the broad covering of the acoustic field by WASNs is to gather the sensors into clusters dominated by a single common source which can then be more efficiently estimated [13].

All these algorithms require the knowledge of either the direction of arrival (DOA) or of the speech activity to compute the filters and are sensitive to signal mismatches [14] or detection errors [6]. Deep learning-based approaches have been proposed to estimate accurately these quantities through the prediction of a time-frequency (TF) mask [15, 16, 17] or of the spectrum of the desired signals [18]. Although often used in a multichannel context, most of these solutions use single-channel data as input of their deep neural networks (DNNs). Multichannel information was first taken into account through spatial features [19], but can also be exploited using the magnitude and phase of several microphones as the input of a convolutional recurrent neural network (CRNN) [20, 21]. Although this yields better results than single-channel prediction, combining all the sensor signals is not scalable and seems suboptimal because of the redundancy of the data. Coping with the redundancy, Perotin et al. [22] combined a single estimate of the source signals with the input mixture and used the resulting tensor to train a long short-term memory (LSTM) recurrent neural network (RNN).

In this paper, we consider a fully connected network with synchronized sensors. This allows for using the MWF-based DANSE algorithm which reported good speech enhancement performances [9]. Following the results shown by Perotin et al. [22], we take profit of the DANSE paradigm [9] by combining at each node one local signal with the estimations of the target signal sent by the other nodes. This uses the multichannel context for the mask estimation but avoids the redundancy brought by the signals of a same node. Additionally, this scheme takes advantage of the internal filter operated in DANSE and reduces the costs in terms of bandwidth and computational power compared to a network combining all the sensor signals.

The paper is organised as follows. The problem formulation and DANSE are described in Section 2. In Section 3 we present

---

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>), and using the EXPLOR centre, hosted by the University of Lorraine.

our solution to estimate the TF masks. The experimental setup is described in Section 4 and results are discussed in Section 5 before we conclude this paper.

## 2. PROBLEM FORMULATION

### 2.1. Signal model

We consider an additive noise model expressed in the short time Fourier transform (STFT) domain as  $y(f, t) = s(f, t) + n(f, t)$  where  $y(f, t)$  is the recorded mixture at frequency index  $f$  and time index  $t$ . The target signal is denoted  $s$  and the noise signal  $n$ . For the sake of conciseness, we will drop the time and frequency indexes  $f$  and  $t$ . The signals are captured by  $M$  microphones and stacked into a vector  $\mathbf{y} = [y_1, \dots, y_M]^T$ . In the following, regular lowercase letters indicate scalars; bold lowercase letters indicate vectors and bold uppercase letters indicate matrices.

### 2.2. Multichannel Wiener filter

The MWF operates on a microphone array or in a fully connected network. It aims at estimating the speech component  $s_i$  of a reference signal at microphone  $i$ . Without loss of generality, we take the reference microphone as  $i = 1$  in the remainder of the paper. The MWF  $\mathbf{w}$  minimises the MSE cost function expressed as follows:

$$J(\mathbf{w}) = \mathbb{E}\{|\mathbf{s}_1 - \mathbf{w}^H \mathbf{y}|^2\}. \quad (1)$$

$\mathbb{E}\{\cdot\}$  is the expectation operator and  $^H$  denotes the Hermitian transpose. The solution to (1) is given by

$$\hat{\mathbf{w}} = \mathbf{R}_{yy}^{-1} \mathbf{R}_{ys} \mathbf{e}_1, \quad (2)$$

with  $\mathbf{R}_{yy} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\}$ ,  $\mathbf{R}_{ys} = \mathbb{E}\{\mathbf{y}\mathbf{s}^H\}$  and  $\mathbf{e}_1 = [1 \ 0 \dots 0]^T$ . Under the assumption that speech and noise are uncorrelated and that the noise is locally stationary,  $\mathbf{R}_{ys} = \mathbf{R}_{ss} = \mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{R}_{yy} - \mathbf{R}_{nn}$  where  $\mathbf{R}_{nn} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}$ . Computing these matrices requires the knowledge of noise-only periods and speech-plus-noise periods. This is typically obtained with a voice activity detector (VAD) [6, 9].

The SDW-MWF provides a trade-off between noise reduction and speech distortion [6]. The filter parameters minimise the cost function

$$J(\mathbf{w}) = \mathbb{E}\{|\mathbf{s}_1 - \mathbf{w}^H \mathbf{s}|^2\} + \mu \mathbb{E}\{|\mathbf{w}^H \mathbf{n}|^2\}, \quad (3)$$

with  $\mu$  the trade-off parameter. The solution to (3) is given by

$$\hat{\mathbf{w}} = (\mathbf{R}_{ss} + \mu \mathbf{R}_{nn})^{-1} \mathbf{R}_{ss} \mathbf{e}_1. \quad (4)$$

If the desired signal comes from a single source, the speech covariance matrix is theoretically of rank 1. Under this assumption, Serizel et al. [23] proposed a rank-1 approximation of  $\mathbf{R}_{ss}$  based on a generalized eigenvalue decomposition (GEVD), delivering a filter that is more robust in low SNR scenarios and provides a stronger noise reduction.

### 2.3. DANSE

In this section, we briefly describe the DANSE algorithm under the assumption that a single target source is present. We consider  $M$  microphones spread over  $K$  nodes, each node  $k$  containing  $M_k$  microphones. The signals of one node  $k$  are stacked in  $\mathbf{y}_k = [y_{k,1}, \dots, y_{k,M_k}]^T$ . As can be seen in (2), the network wide MWF should be computed from all signals of the network, which

can result in high bandwidth and computational costs. In DANSE, at node  $j$ , only a single compressed signal  $z_j$  is sent to the other nodes. So a node  $k$  disposes of  $M_k + K - 1$  signals, stacked in  $\tilde{\mathbf{y}}_k = [\mathbf{y}_k^T, \mathbf{z}_{-k}^T]^T$ , where  $\mathbf{z}_{-k}$  is a column vector gathering the compressed signals coming from the other nodes  $j \neq k$ . Replacing  $\mathbf{y}$  by  $\tilde{\mathbf{y}}_k$  and solving (3) yields the DANSE solution to the SDW-MWF:

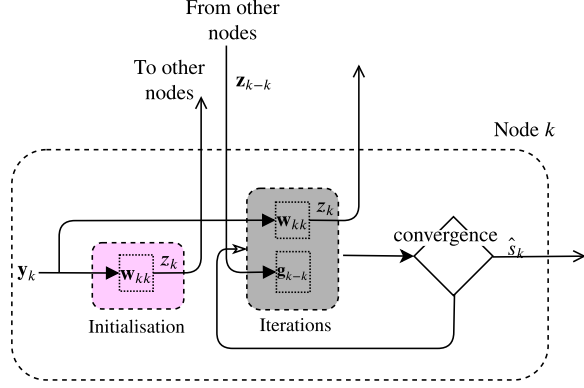
$$\tilde{\mathbf{w}}_k = (\mathbf{R}_{ss,k} + \mu \mathbf{R}_{nn,k})^{-1} \mathbf{R}_{ss,k} \mathbf{e}_1, \quad (5)$$

where  $\tilde{\mathbf{w}}_k$ , the filter at node  $k$ , can be decomposed into two filters as  $\tilde{\mathbf{w}}_k = [\mathbf{w}_{kk}^T, \mathbf{g}_{k-k}^T]^T$ . The first filter  $\mathbf{w}_{kk}^T$  is applied on the local signal and  $\mathbf{g}_{k-k}$  is applied on the compressed ones sent from the other nodes. The covariance matrices  $\mathbf{R}_{ss,k}$  and  $\mathbf{R}_{nn,k}$  are computed from the speech and noise components of  $\tilde{\mathbf{y}}_k$ . The compressed signal  $z_k$  is computed as  $z_k = \mathbf{w}_{kk}^H \mathbf{y}_k$ . Bertrand and Moonen proved that this solution converges to the MWF solution with  $\mu = 1$ , while dividing the bandwidth load by a factor  $M_k$  at each node [9]. In this paper, we will focus on the batch-mode algorithm where the speech and noise statistics are computed based on the whole signal in order to focus on the interaction between the mask estimated by the DNN and the MWF filters. This delivers a fully spatial processing since the filters are time-invariant, while delaying the study of the convergence hyperparameters to future work.

## 3. DEEP NEURAL NETWORK BASED DISTRIBUTED MULTICHANNEL WIENER FILTER

Heymann et al. predicted TF masks out of a single signal of the microphone array [16]. Perotin et al. [22] or Chakrabarty and Habets [21] included several other signals to improve the speech recognition or speech enhancement performances. We propose to extend these scenarios to the multi-node context of DANSE. In DANSE, at node  $k$ , a single VAD is used to estimate the source and noise statistics required for both filters  $\mathbf{w}_{kk}$  and  $\mathbf{w}_k$ . The first part of our contribution is to replace the VAD by a TF mask predicted by a DNN. Besides, since the compressed signals  $z_k$  are sent from one node to the others, we also examine the option of exploiting this extra source of information by using it for the mask prediction. The schematic principle of DANSE is depicted in Figure 1. As it can be seen, an initialisation phase is required to first compute the signal  $z_k$ . Bertrand et al. initialise them randomly, but we propose to do this with a first neural network. The second stage of DANSE is represented in the greyed box in Figure 1 and expended in Figure 2. Our second contribution is highlighted with the red arrow. It is to exploit the presence of  $\mathbf{z}_{-k}$  at one node to better predict the masks with the DNN. Several iterations are necessary for the filter  $\mathbf{w}_{kk}$  to converge to the solution (4). In DANSE, iterations are done at every time step. As we developed an offline batch-mode algorithm, we stopped the processing after the first iteration. To analyse the effectiveness of combining  $\mathbf{z}_{-k}$  with a reference signal to predict the mask, we compare our solution with a single-channel prediction, where the masks required for both initialising and iterated stages are predicted by a single-channel model seeing only the local signal  $y_{k,1}$ .

We compare two different architectures for each of these schemes. The first architecture is a bidirectional LSTM introduced by Heymann et al. [16]. When a RNN was used with additional input data, the additional channels  $\mathbf{z}_{-k}$  were stacked over the frequency axis [22]. Although this delivered improved performance compared to single-channel version (see also Section 5), stacking it over the frequency axis is not efficient as many connections are used to represent relations between TF bins that might not be related. That is



**Fig. 1.** Block diagram of DANSE principle. Plain arrows represent vectors, simple ones represent scalars.

why we propose a CRNN architecture which is more appropriate to process multichannel data. Each acoustic channel is then considered as a convolutional channel.

During the training, in order to take into account the spectral shape of the speech, we weight the MSE loss between the predicted mask  $\hat{\mathbf{m}}$  and the ground truth mask  $\mathbf{m}$  by the STFT frame of the input  $\mathbf{y}$ , corresponding to the predicted frame. Both models are thus trained to minimise the cost function

$$\mathcal{L}(\mathbf{m}, \hat{\mathbf{m}}) = E\{|\mathbf{m} - \hat{\mathbf{m}}| \cdot \mathbf{y}^2\},$$

where  $E\{\cdot\}$  represents the empirical mean.

Lastly, since the filter  $\mathbf{w}_{kk}$  is also applied on  $\mathbf{z}_{-k}$ , we used the GEVD of the covariance matrices to compute the MWF of equation (4). Contrary to equation (2), this does not explicitly take the first microphone as a reference. It also allows for a higher importance assigned to the compressed signals, which is desirable since they are pre-filtered with potentially higher signal to noise ratios (SNRs) than the local signals.

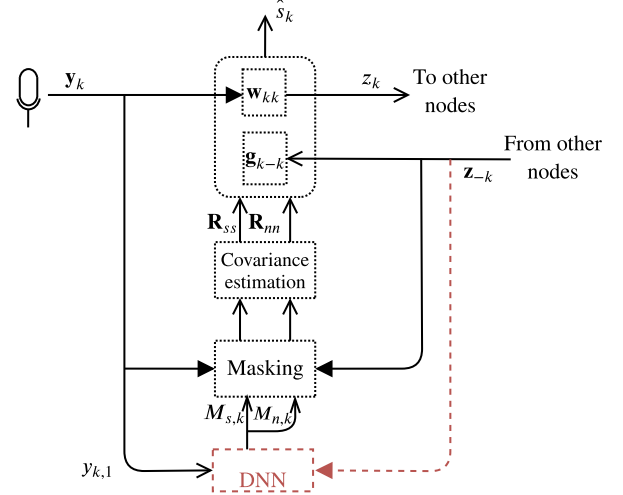
## 4. EXPERIMENTAL SETUP

### 4.1. Dataset

Training as well as test data was generated by convolving clean speech and noise signals with simulated room impulse responses (RIRs), and then by mixing the convolved signals at a specific SNR. The anechoic speech material was taken from the clean subset of LibriSpeech [24]. The RIRs were obtained with the Matlab toolbox *Roomsimove*<sup>1</sup> simulating shoebox-like rooms.

In the training set, the length of the room drawn uniformly as  $l \in [3, 8]$  m, the width as  $w \in [3, 5]$  m, the height as  $h \in [2, 3]$  m. Two nodes of four microphones each recorded the acoustic scene. The distance between the nodes was set to 1 m, the microphones 10 cm away from the node centre. Each node was at least 1 m away from the closest wall. One source of noise and one of speech were placed at 2.5 m from the array centre. Both sources had an angular distance  $\alpha \in [25, 90]^\circ$  relative to the array centre. Microphones as well as sources were at the constant height of 1.5 m. The SNR was drawn uniformly between  $-5$  dB and  $+15$  dB. The noise was white noise modulated in the spectral domain by the long term spectrum of speech. We generated 10,000 files of 10 seconds each, corresponding to about 25 hours of training material.

<sup>1</sup>homepages.loria.fr/evincent/software/Roomsimove\_1.4.zip



**Fig. 2.** Expansion of the iterated step in Figure 1. Red arrows are the modifications brought to DANSE. Plain arrows represent vectors, simple ones represent scalars.

The test configurations is the same as the training configuration but with restricted values for several parameters. The length of the room was randomly selected among  $l \in \llbracket 3, 8 \rrbracket$  m, the width as  $w \in \llbracket 3, 5 \rrbracket$  m, the height was set to  $h = 2.5$  m. The angular distance  $\alpha$  between the sources was randomly selected in  $\alpha = \{25, 45, 90\}^\circ$ . The noise was a random part of the third CHiME challenge dataset [25] in the cafeteria or pedestrian environment. We generated 1,000 files representing about 2 hours of test material.

### 4.2. Setup

All the data was sampled at 16 kHz. The STFT was computed with an FFT-length of 512 samples (32 ms), 50% overlap and a Hanning window. After several trials, the only modification we brought to Heymann's setup was that the input data was made of sequences of 21 STFT frames and that the mask for the middle one had to be predicted. This brought better performances than when the mask for whole STFT segments had to be predicted at once.

Our CRNN model is composed of three convolutional layers with 32, 64 and 64 filters respectively. They all have  $3 \times 3$  kernels, with stride  $1 \times 1$  and ReLU activation functions. Each convolutional layer is followed by a batch normalization over the frequency axis and a maximum pooling layer of size  $4 \times 1$  (along the frequency axis). The recurrent part of the network is a gated recurrent unit (GRU) layer with 256 units, and the last layer is a fully connected layer with a sigmoid activation function. The CRNN was also given 21 frames at the input. Both networks had to predict the middle frame and were trained with the RMSprop optimizer [26].

## 5. RESULTS

We evaluate the speech enhancement performances based on the source to artifacts ratio (SAR), source to interferences ratio (SIR) and source to distortion ratio (SDR) [27] computed with the

*mir\_eval*<sup>2</sup> toolbox. The performance reported corresponds to the mean over the 1,000 test samples of the objective measures computed at the node with the best input SNR. We also report the 95% confidence interval.

The GEVD filter does not explicitly take one sensor signal as the reference signal to minimise the cost function, but a projection of the input signals into the space spanned by the common eigenvectors of the covariance matrices. Because of that, the objective measures computed with respect to the convolved signals did not give results that were coherent with perceptual listening tests performed internally on random samples. Indeed, differences between the enhanced signal and the reference signal are interpreted as artefacts whereas they are due to the decomposition of the input signals into the eigenvalue space of the covariance matrices. Therefore, we compute the objective measures using the dry (source) signals as reference signals. This decreases the SAR because the reverberation is then considered as an artefact but the comparison between methods is valid and correlates more with the perceptual listening tests.

We present the objective metrics for different approaches in Table 1. In this table, single node filters are referred to as MWF (upper part of the table) and distributed filters as DANSE (lower part of the table). For each filter, the architecture used to obtain the masks is indicated between parenthesis. RNN refers to Heymann’s architecture and CRNN to the network introduced in Section 4.2. The subscript of the network architecture indicates the channels considered at the input. The results obtained with the single-channel DNN models are denoted with “SC”. When the compressed signals  $\mathbf{z}_{-k}$  were additionally used as input to the DNN to predict the mask of the second filtering stage, models are denoted with “MC” (this corresponds to the last two rows of the table). Additionally, we report in Table 2 the number of trainable parameters of each model.

### 5.1. Oracle performances

The VAD gives information about the speech-plus-noise and noise-only periods in a time-wise manner only, whereas a mask gives spectral information that enables a finer estimation of the speech and noise covariance matrices. This additional information translates in an improvement of the speech enhancement performances with both types of filters (MWF and DANSE). In the following section, we analyse whether this conclusion still holds when the masks are predicted by a neural network.

### 5.2. Performances with predicted masks

First, replacing the oracle VAD by a predicted mask brings significant improvement in terms of all objective measures. This confirms the idea that TF masks are better activity detectors than VADs, even oracle one. Second, the objective measures corresponding to the output signals of DANSE filters are always better than those of the MWF filters. This confirms the benefit of using the DANSE algorithm. Although these differences are not high, increasing the number of nodes and the distance between them might enhance the utility of the distributed method.

Using a CRNN structure over a RNN one is not revealed when single-channel data is used. Indeed, the objective measures of  $\text{RNN}_{\text{SC}}$  and  $\text{CRNN}_{\text{SC}}$  match in all points. In the multichannel case, the performance of the RNN-based approach does not increase. This tends to confirm that the RNN is not able to efficiently exploit multichannel information. This is probably not because the architecture is less suited, but because stacking the signals on the frequency axis

(dB)	SAR	SIR	SDR
MWF (oracle VAD)	2.4±0.3	24.7±0.3	2.3±0.3
MWF (oracle mask)	4.0±0.3	26.7±0.3	3.9±0.3
MWF (RNN)	3.4±0.3	25.1±0.4	3.3±0.3
MWF (CRNN)	3.3±0.3	25.1±0.4	3.2±0.3
DANSE (oracle VAD)	2.6±0.3	25.2±0.3	2.6±0.3
DANSE (oracle mask)	4.8±0.3	27.6±0.3	4.8±0.3
DANSE ( $\text{RNN}_{\text{SC}}$ )	4.0±0.3	26.0±0.4	4.0±0.3
DANSE ( $\text{CRNN}_{\text{SC}}$ )	4.0±0.3	26.0±0.4	4.0±0.3
DANSE ( $\text{RNN}_{\text{MC}}$ )	4.1±0.3	26.1±0.4	4.0±0.3
DANSE ( $\text{CRNN}_{\text{MC}}$ )	4.7±0.3	27.4±0.3	4.6±0.3

**Table 1.** Speech enhancement results in dB with oracle activity detectors and predicted ones.

Model	Number of parameters
$\text{RNN}_{\text{SC}}$	1,717,773
$\text{CRNN}_{\text{SC}}$	911,109
$\text{RNN}_{\text{MC}}$	2,244,109
$\text{CRNN}_{\text{MC}}$	911,397

**Table 2.** Number of trainable parameters of the neural networks.

is not meaningful. In addition, as shown in Table 2, the number of parameters of the RNN almost doubles when a second signal is used, whereas it barely increases for the CRNN. This is due to the convolutional layers of the CRNN which can much more efficiently process multichannel data than recurrent layers.

The CRNN solution can exploit the multichannel inputs efficiently and the performance increases for all metrics. The biggest improvement is done for the SIR. An explanation can be that one of the main difficulty for the models is to predict noise-only regions, because of people talking in the noise CHiME database. Since the compressed signal  $\mathbf{z}_{-k}$  are pre-filtered, they contain less noise and they are less ambiguous. This makes it easier for the model to recognize noise-only regions, without degrading its predictions of speech-plus-noise regions.

## 6. CONCLUSION AND FUTURE WORK

We introduced a efficient way of estimating masks in a multi-node context. We developed multichannel models combining an estimation of the target signals sent by the other nodes with a local sensor. This proved to better predict TF masks, which led to higher speech enhancement performances that outperformed the results obtained with an oracle VAD. A CRNN architecture was compared to a RNN one and the CRNN architecture could exploit much better the multichannel information. In addition, the RNN structure is limited by its number of parameters, especially if the number of nodes had to increase. In such scenarios, the difference between single-channel and multichannel models performances might be even more important but this still has to be inquired. To attain performances closer to the oracle ones, several options are possible. First, the rather simple architectures that were used could be replaced by state-of-the art architectures. Besides, given the increase in performance when the target estimation is given, it would also be interesting to additionally give the noise estimation at the input of the models. Lastly, an adaptive version seems necessary to dedicate this algorithm to real-time applications.

<sup>2</sup>[https://github.com/craffel/mir\\_eval/](https://github.com/craffel/mir_eval/)

## 7. REFERENCES

- [1] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.
- [2] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE GlobalSIP*. IEEE, 2014, pp. 577–581.
- [3] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.
- [4] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio source separation and speech enhancement*, John Wiley edition, 2018.
- [5] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.
- [6] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, 2007.
- [7] Alexander Bertrand, Simon Doclo, Sharon Gannot, Nobutaka Ono, and Toon van Waterschoot, "Special issue on wireless acoustic sensor networks and ad hoc microphone arrays," *Signal Processing*, vol. 107, no. C, pp. 1–3, 2015.
- [8] A. Bertrand, J. Callebaut, and M. Moonen, "Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks," in *Proc. of IWAENC*, 2010.
- [9] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks - Part I: Sequential node updating," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5277–5291, 2010.
- [10] Y. Zeng and R. C. Hendriks, "Distributed estimation of the inverse of the correlation matrix for privacy preserving beamforming," *Signal Processing*, vol. 107, pp. 109–122, 2015.
- [11] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed mvdr beamforming for (wireless) microphone networks using message passing," in *IWAENC*, 2012, pp. 1–4.
- [12] M. O'Connor and W. B. Kleijn, "Diffusion-based distributed MVDR beamformer," in *IEEE Proc. of ICASSP*, 2014, pp. 810–814.
- [13] S. Gergen, R. Martin, and N. Madhu, "Source separation by feature-based clustering of microphones in ad hoc arrays," *IWAENC*, pp. 530–534, 2018.
- [14] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 313–324, 2003.
- [15] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *IEEE ICASSP*, pp. 7092–7096, 2013.
- [16] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE ICASSP*, 2016, vol. 2016-May, pp. 196–200.
- [17] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 241–245.
- [18] A.A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1652–1664, 2016.
- [19] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.
- [20] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *EUSIPCO*, Sep. 2018, pp. 1462–1466.
- [21] S. Chakrabarty and E. A. P. Habets, "Time-Frequency Masking Based Online Multi-Channel Speech Enhancement With Convolutional Recurrent Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 1–1, 2019.
- [22] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *IEEE Proc. of ICASSP*, 2018, pp. 36–40.
- [23] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.
- [24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE Proc. of ICASSP*, 2015, pp. 5206–5210.
- [25] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE ASRU*, December 2015, pp. 504–511.
- [26] G. Hinton, N. Srivastava, and K. Swersky, "COURSERA: Neural networks for machine learning – lecture 6a," 2012. Online available at [http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf).
- [27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.