Distributed speech separation in spatially unconstrained microphone arrays

Nicolas Furnon[1], Romain Serizel[1], Irina Illina[1], Slim Essid[2]

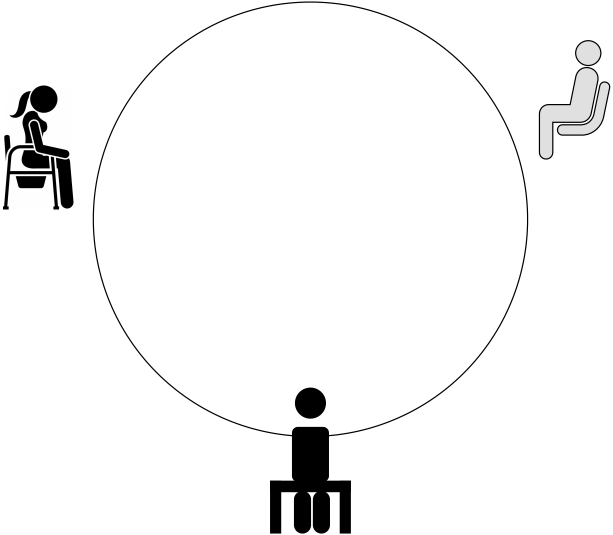
[1]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
{firstname.lastname}@loria.fr
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
slim.essid@telecom-paris.fr

ICASSP 2021

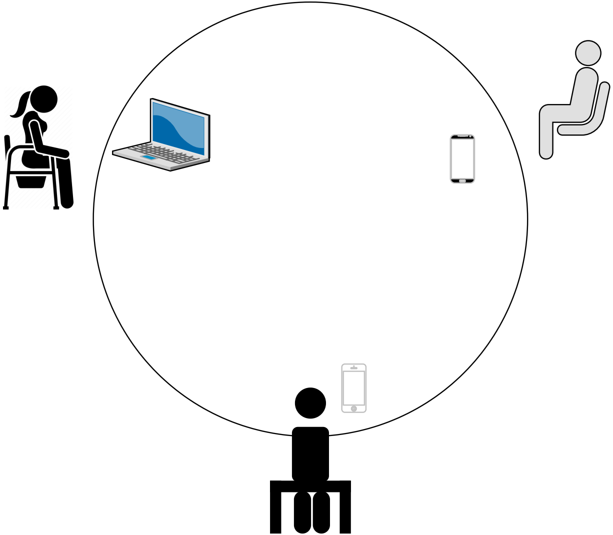
Some icons downloaded from http://pinclipart.com, https://iconfinder.com

Structure

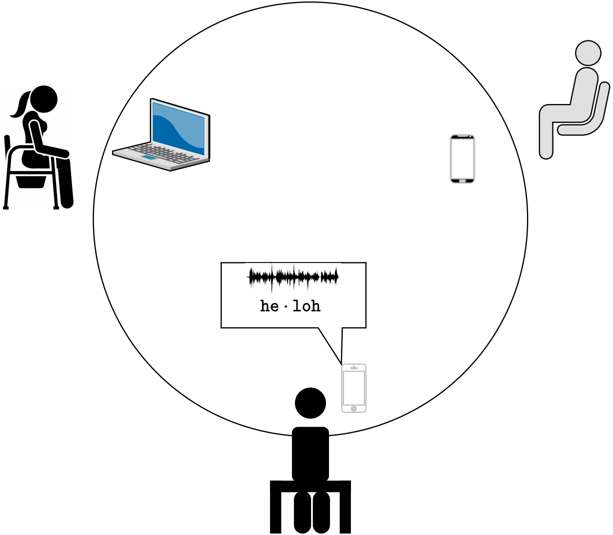
A typical meeting scenario



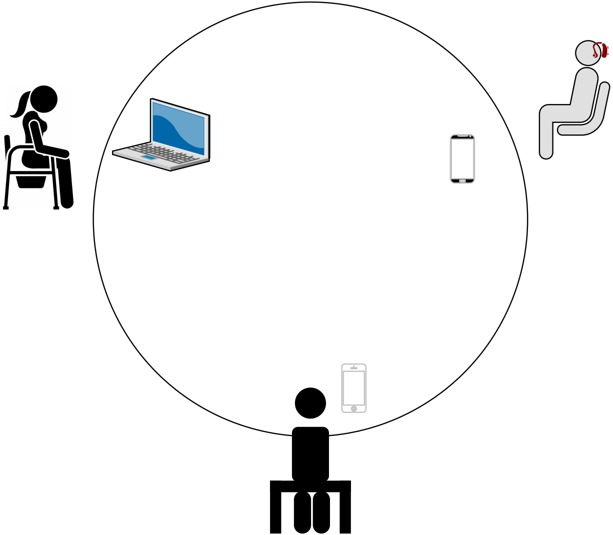
A typical meeting scenario



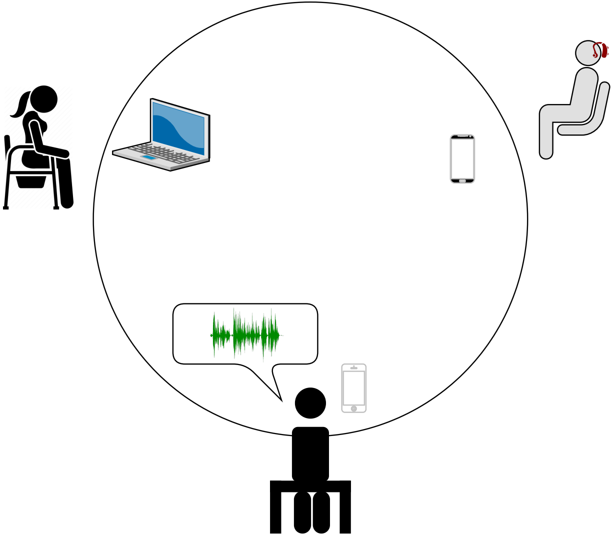
A typical meeting scenario



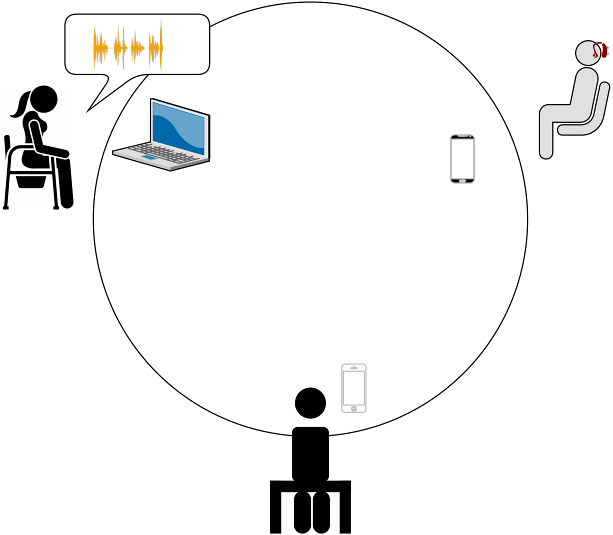
A typical meeting scenario



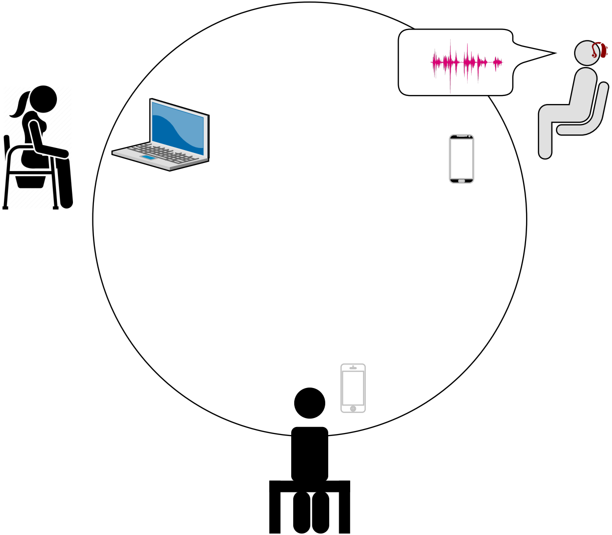
Source separation



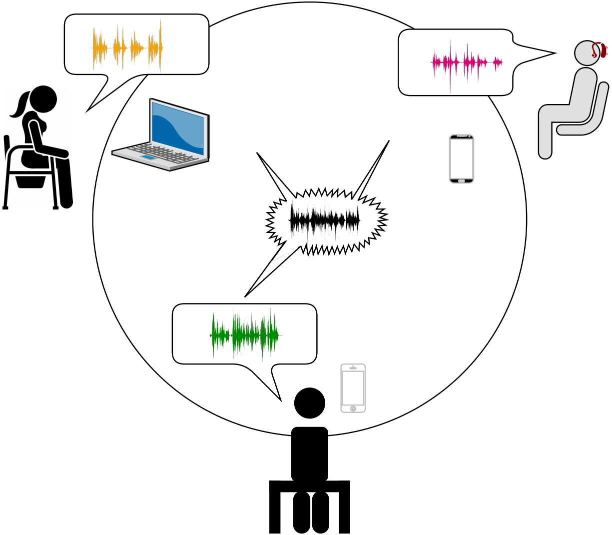
Source separation



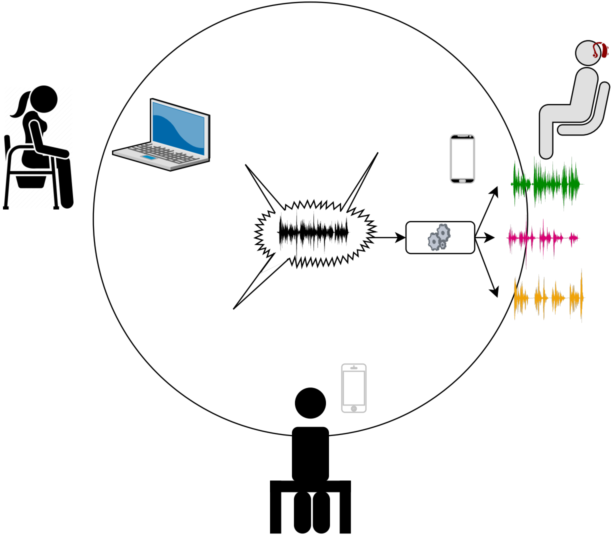
Source separation



Source separation



Source separation



State-of-the-art

Reference	👍	👎
[1, 2, 3, 4]	Efficient performance	• Unrealistic conditions • No spatial information
[5, 6] [7]	• Multichannel information • Reverberant conditions	Complex NNs

[1] Y. Luo and N. Mesgarani, *Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation* (2019)
[2] L. Zhang et al. *FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks* (2020)
[3] N. Zeghidour and D. Grangier, *Wavesplit: End-to-end speech separation by speaker clustering* (2020)
[4] J. Chen et al. *Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation* (2020)
[5] R. Gu et al. *End-to-end multi-channel speech separation* (2019)
[6] D. Wang et al. *Neural speech separation using spatially distributed microphones* (2020)
[7] M. Delfarah and D. Wang, *Deep learning for talker-dependent reverberant speaker separation: An empirical study* (2019)

Contribution

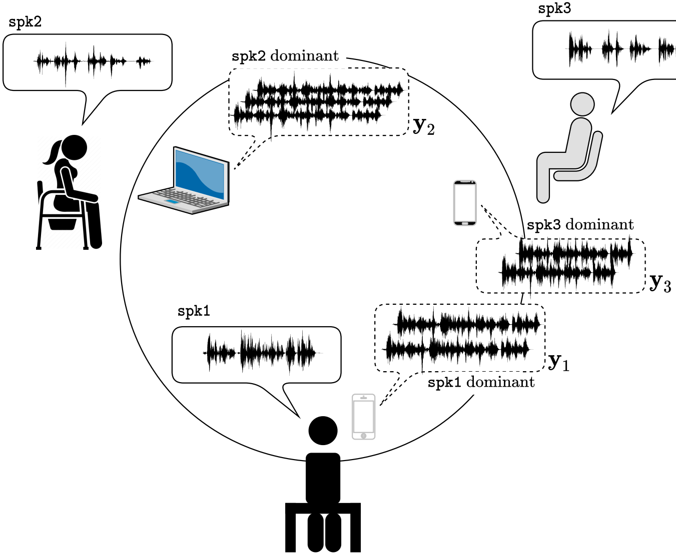
Distributed speech separation

"Tango" [8]: a two-step source separation algorithm based on a distributed adaptive node-specific signal estimation (DANSE) [9]

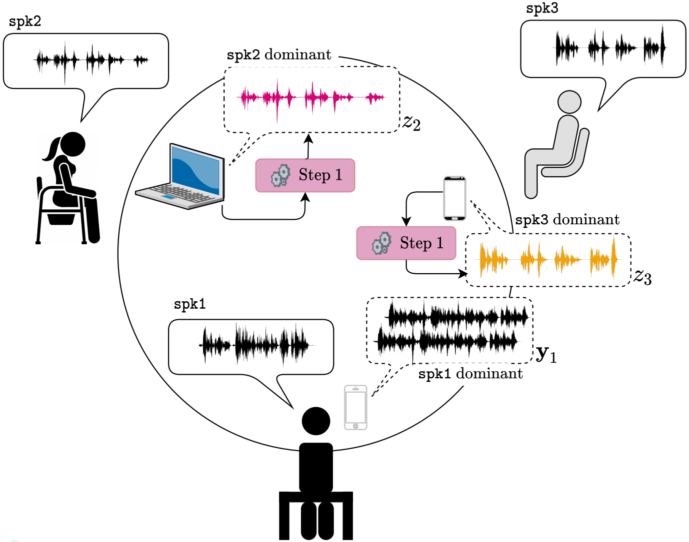
[8] N. Furnon et al. *DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays* (2021)

[9] A. Bertrand and M. Moonen *Distributed adaptive node-specific signal estimation in fully connected sensor networks — Part I: Sequential node updating* (2010)

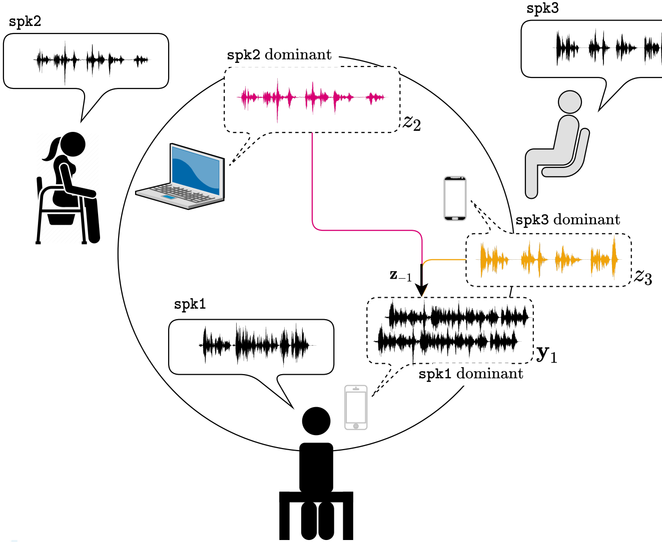
Tango – Overview



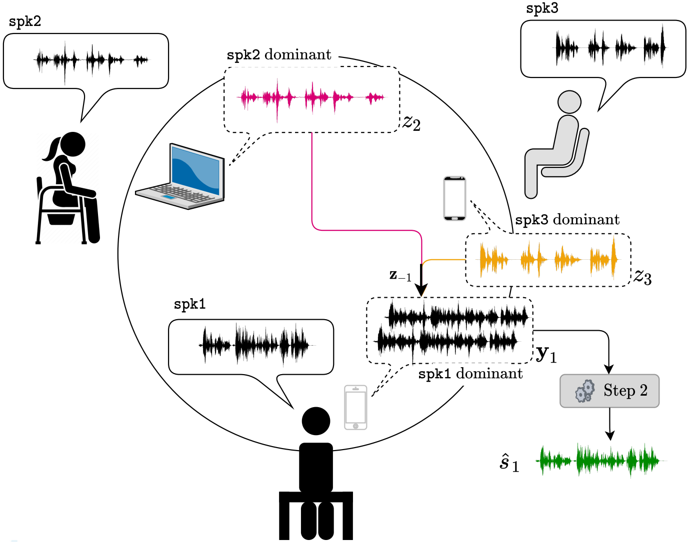
Tango – Overview



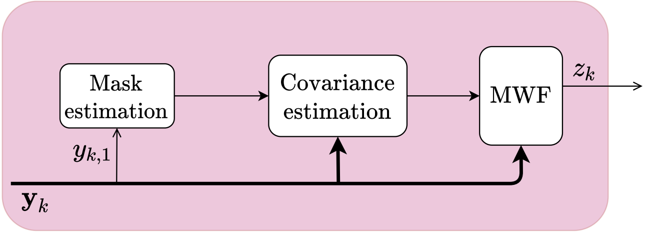
Tango — Overview



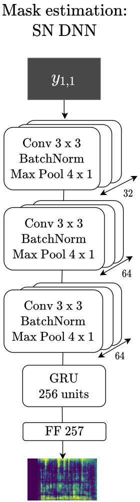
Tango – Overview



Tango – Step 1

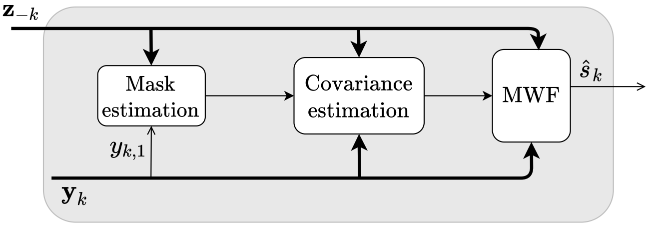


Mask estimation
Covariance estimation
MWF
z_k
$y_{k,1}$
$\mathbf{y}_k$

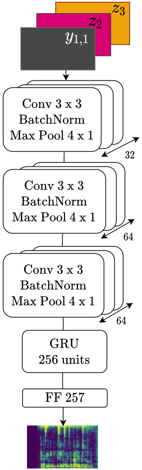


Mask estimation:
SN DNN
$y_{1,1}$
Conv 3 x 3
BatchNorm
Max Pool 4 x 1
32
Conv 3 x 3
BatchNorm
Max Pool 4 x 1
64
Conv 3 x 3
BatchNorm
Max Pool 4 x 1
64
GRU
256 units
FF 257

Tango – Step 2



Mask estimation:
MN DNN



Tango

Advantages

- Usage of a priori knowledge (local SNR): each node estimates and sends a different source
- Spatial information sent as pre-filtered estimates
- Exploitation of spatial information
- Distributed processing

Experiments

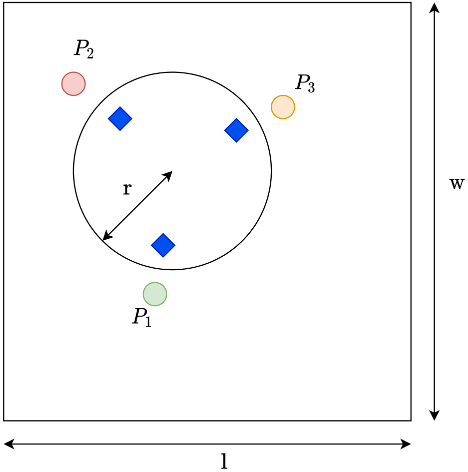
Parameters

1 node = 4 microphones

$$N = \{2, 3, 4\} \text{ sources}$$
$$K = \{2, 3, 4\} \text{ nodes}$$

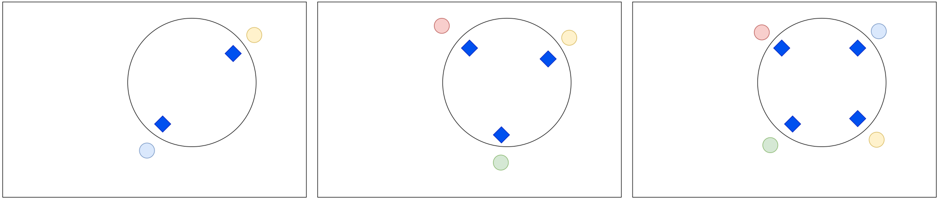
$$0.3 \text{ m} \leq r \leq 2.5 \text{ m}$$
$$3 \text{ m} \leq w \leq 7 \text{ m}$$
$$3 \text{ m} \leq l \leq 9 \text{ m}$$
$$150 \text{ ms} \leq T_{60} \leq 400 \text{ ms}$$

$$P_1 = P_2 = P_3 \text{ (dry source levels)}$$



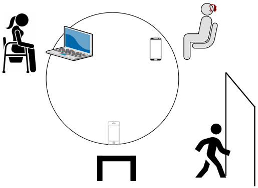
The code to reproduce the dataset is available at `https://github.com/nfurnon/disco/tree/master/dataset_generation`

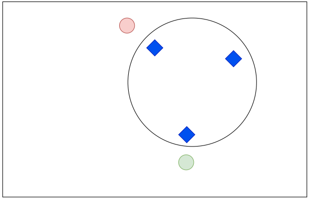
Balanced cases ($K = N$)



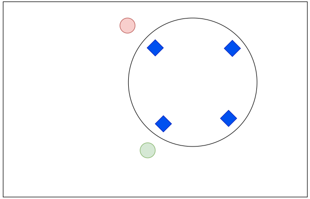
$N = K = 2$ $N = K = 3$ $N = K = 4$

Over-determined cases ($K > N$)

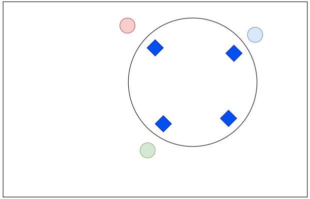




$K = 3, N = 2$

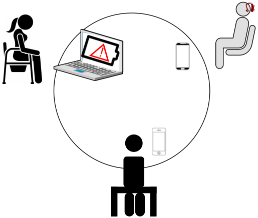


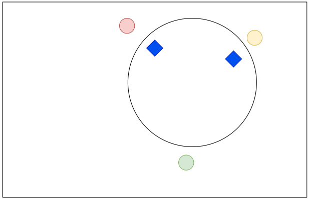
$K = 4, N = 2$



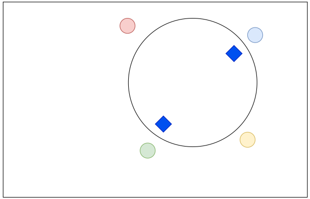
$K = 4, N = 3$

Under-determined cases ($K < N$)

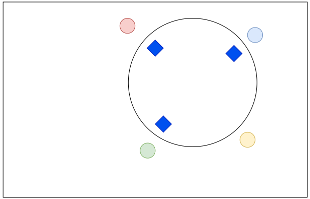




$K = 2, N = 3$



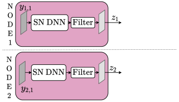
$K = 2, N = 4$



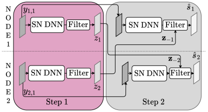
$K = 3, N = 4$

Compared methods

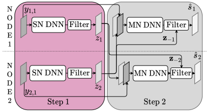
1. MWF



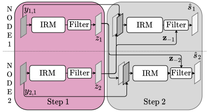
2. SN



3. MN

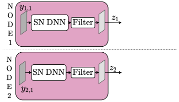


4. IRM

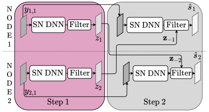


Compared methods

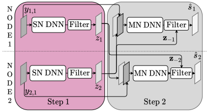
1. MWF



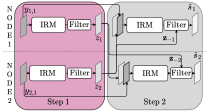
2. SN



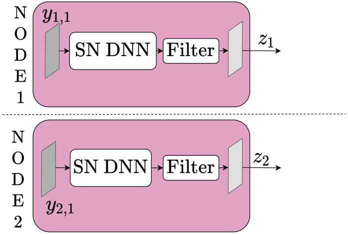
3. MN



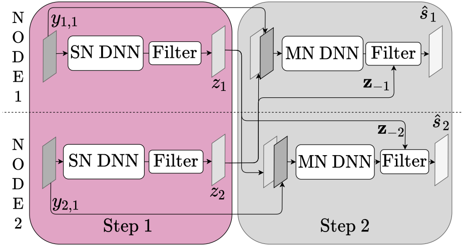
4. IRM



Compared methods



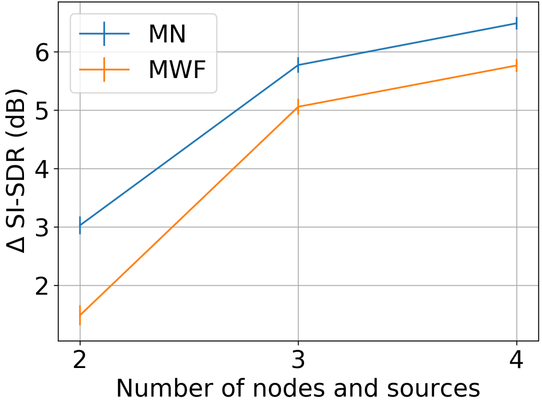
MWF



MN

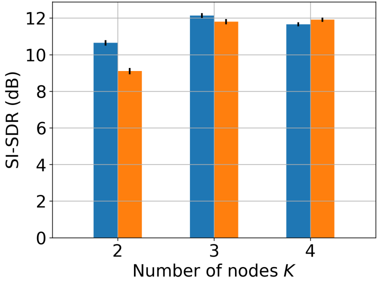
Results

Balanced cases ($K = N$)

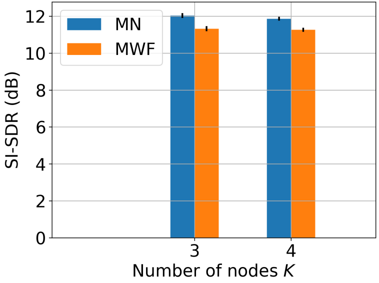


- MN > MWF: **Compressed signals useful also for the mask estimation**
- Increasing ΔSI-SDR with increasing K, N: **Robustness to spatial diversity**

Over-determined cases ($K \geq N$)



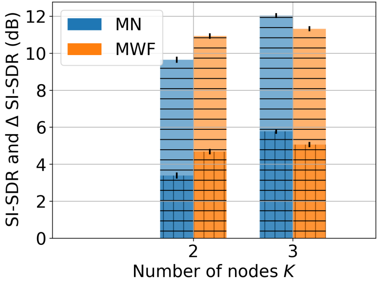
$N = 2$



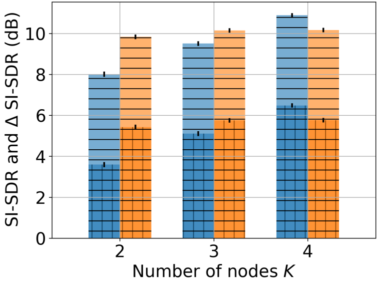
$N = 3$

- MN > MWF almost always **Robust to training/testing mismatches**
- Exception at $N = 2, K = 4$: Compressed signals are silent
- Increased performance from $K = 2 \rightarrow K = 3$: **Better to train on harder conditions**

Under-determined cases ($K \leq N$)



$N = 3$



$N = 4$

- MN < MWF: **MN CRNN fails in mismatched conditions**
- Need for a dedicated strategy [10, 11]

[10] K. Kinoshita et al. *Listening to each speaker one by one with recurrent selective hearing networks* (2018).
[11] N. Turpault et al. *Improving sound event detection in domestic environments using sound separation* (2020).

Conclusion

Tango: a distributed processing for source separation

- Can process spatial information
- Evaluated on realistic meeting scenarios
- Improves performance when the number of nodes (and sources) increases
- Restricted to equally-determined or over-determined cases

Thank you for your attention

nicolas.furnon@loria.fr

References I

[1] Yi Luo and Nima Mesgarani,
"Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,"
IEEE/ACM transactions on audio, speech, and language processing, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] Liwen Zhang, Ziqiang Shi, Jiqing Han, Anyan Shi, and Ding Ma,
"FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks,"
in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 653–665.

[3] Neil Zeghidour and David Grangier,
"Wavesplit: End-to-end speech separation by speaker clustering,"
arXiv preprint arXiv:2002.08933, 2020.

[4] Jingjing Chen, Qirong Mao, and Dong Liu,
"Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,"
arXiv preprint arXiv:2007.13975, 2020.

[5] Rongzhi Gu, Jian Wu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu,
"End-to-end multi-channel speech separation,"
arXiv preprint arXiv:1905.06286, 2019.

[6] Dongmei Wang, Zhuo Chen, and Takuya Yoshioka,
"Neural speech separation using spatially distributed microphones,"
arXiv preprint arXiv:2004.13670, 2020.

[7] Masood Delfarah and DeLiang Wang,
"Deep learning for talker-dependent reverberant speaker separation: An empirical study,"
IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1839–1848, 2019.

References II

[8] Nicolas Furnon, Romain Serizel, Irina Illina, and Slim Essid,
"DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays,"
submitted to IEEE/ACM Transactions on Audio Speech and Language Processing, 2020.

[9] Alexander Bertrand and Marc Moonen,
"Distributed adaptive node-specific signal estimation in fully connected sensor networks — Part I: Sequential node updating," Oct 2010.

[10] Keisuke Kinoshita, Lukas Drude, Marc Delcroix, and Tomohiro Nakatani,
"Listening to each speaker one by one with recurrent selective hearing networks,"
in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5064–5068.

[11] Nicolas Turpault, Scott Wisdom, Hakan Erdogan, John Hershey, Romain Serizel, Eduardo Fonseca, Prem Seetharaman, and Justin Salamon,
"Improving sound event detection in domestic environments using sound separation,"
arXiv preprint arXiv:2007.03932, 2020.

Loria