DNN-based distributed multichannel mask estimation for speech enhancement in microphone arrays

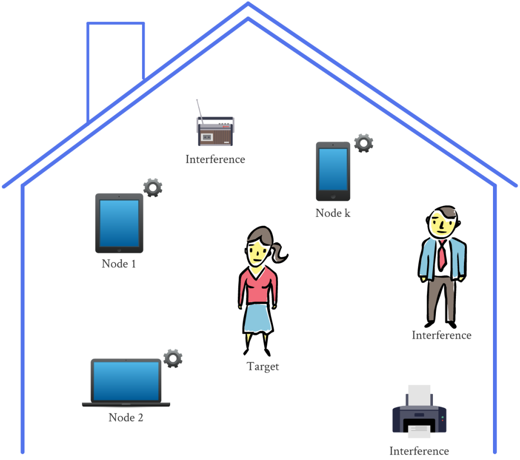
Nicolas Furnon[1], Romain Serizel[1], Irina Illina[1], Slim Essid[2]

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

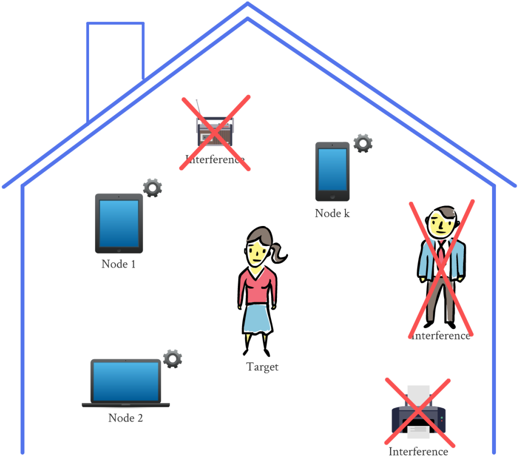
ICASSP 2020

Structure

Speech enhancement in ad-hoc microphone arrays



Speech enhancement in ad-hoc microphone arrays



Speech enhancement in ad-hoc microphone arrays

Advantages

- Flexible unconstrained geometry and usage
- Larger area coverage

Challenges

- **Distributed processing**
- Synchronization and calibration among nodes

Speech enhancement in ad-hoc microphone arrays

Distribute the processing for scalable, power-limited solutions

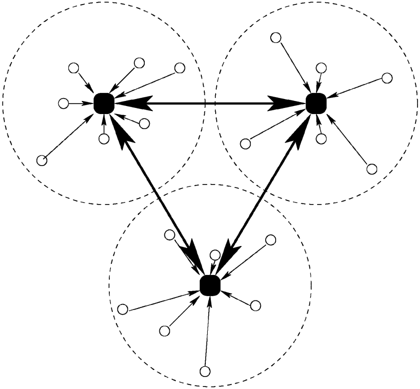
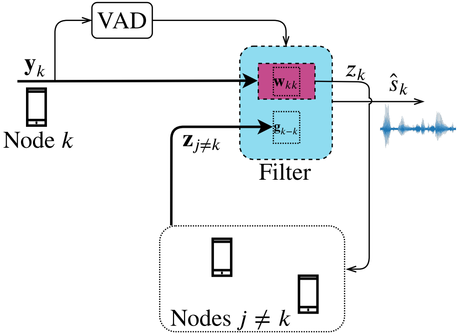


Figure from [Bertrand and Moonen, 2010]

DANSE algorithm [Bertrand and Moonen, 2010]



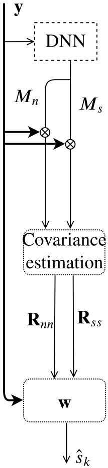
$$\hat{s}_k = \mathbf{w}_{kk}^H \cdot \mathbf{y}_k + \mathbf{g}_{k-k}^H \cdot \mathbf{z}_{j\neq k}$$

$$z_k = \mathbf{w}_{kk}^H \cdot \mathbf{y}_k$$

DNN-based multichannel speech enhancement

Use DNN to predict:

- **TF-masks** [Heymann et al., 2016]
- Clean spectrograms [Nugraha et al., 2016]
- Beamformer coefficients [Pfeifenberger et al., 2019]

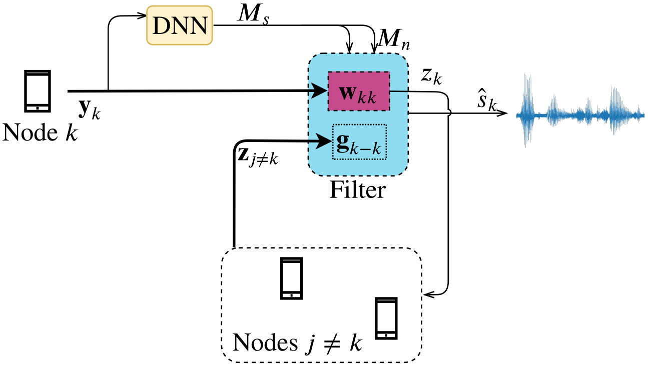


Proposed solution

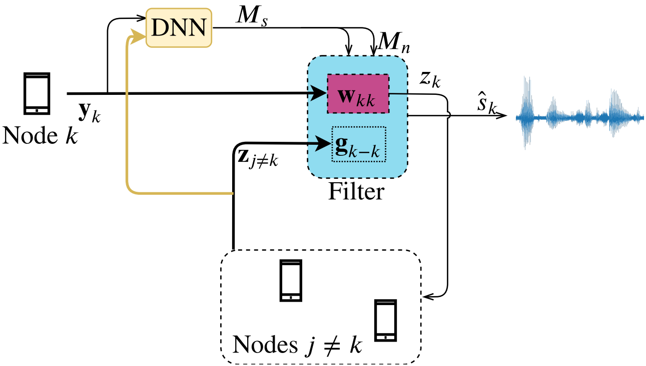
Bridge the gap between distributed solutions and DNN-based solutions

- In DANSE, replace the VAD by a DNN-predicted TF-mask
- Exploit the multichannel information

DNN-based mask estimation



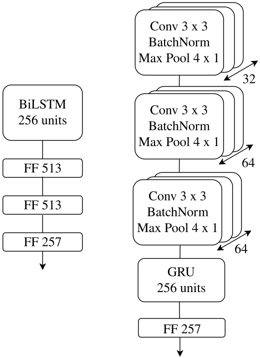
Exploitation of the multi-node context



Use the compressed signals to better predict the mask

Comparison of the DNN architectures

RNN Vs CRNN

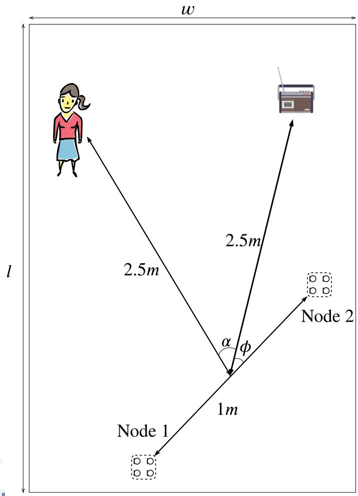


RNN [Heymann et al., 2016]

+ Process the temporal information with the recurrent layers

CRNN [Perotin et al., 2018]

+ Scalable to an increase of input channels
+ Efficient processing of multichannel input



Acoustic scenario

Train dataset: 10.000 signals

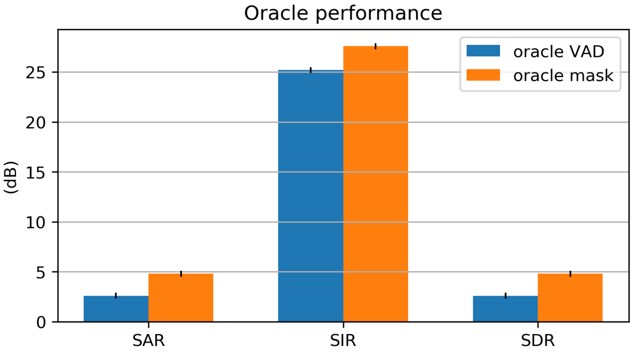
- $\alpha \in [25, 90]^\circ$
  $\phi \in [-180, 180]^\circ$
- $3 \times 3 \times 2 \leq l \times w \times h \leq 8 \times 5 \times 3\,\mathrm{m}$
- $T_{60} \in [300, 600]\,\mathrm{ms}$
- $SNR \in [-5, 15]\,\mathrm{dB}$
- Speech : LibriSpeech
- Noise: SSN

Test dataset: 1.000 signals

Same as train dataset but with restricted values for the parameters.

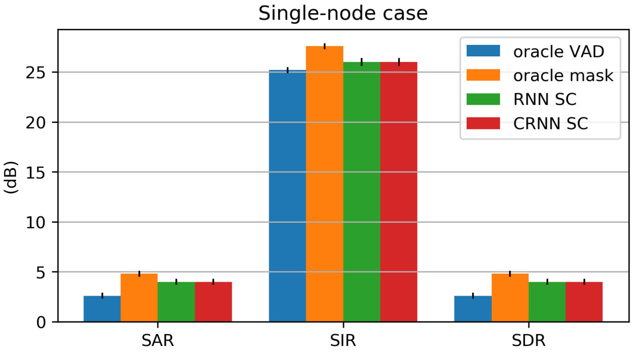
Main difference: noise from CHiME 3

Performance with oracle activity detectors



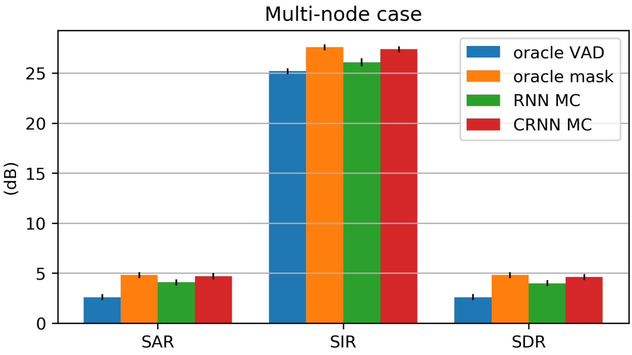
- Useful to use a mask instead of a VAD

RNN Vs CRNN: single-node case



- Similar performances as with an oracle VAD
- No much difference between RNN and CRNN

RNN Vs CRNN: multi-node case



- No improvement with the RNN
- Significant improvement with the CRNN

Conclusion

Conclusions

- First DNN-based distributed speech enhancement algorithm
- Exploitation of the multi-node context for a more accurate mask estimation

Perspectives

- Generalization to scenarios with a higher number of nodes and varied signals
- Better exploitation of the information coming from the other nodes (e.g. exploit SNR diversity)

References

[Bertrand and Moonen, 2010] Bertrand, A. and Moonen, M. (2010).
Distributed adaptive node-specific signal estimation in fully connected sensor networks - Part I: Sequential node updating.
IEEE Transactions on Signal Processing, 58(10):5277–5291.

[Heymann et al., 2016] Heymann, J., Drude, L., and Haeb-Umbach, R. (2016).
Neural network based spectral mask estimation for acoustic beamforming.
In *IEEE ICASSP*, volume 2016-May, pages 196–200.

[Nugraha et al., 2016] Nugraha, A., Liutkus, A., and Vincent, E. (2016).
Multichannel audio source separation with deep neural networks.
IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(10):1652–1664.

[Perotin et al., 2018] Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018).
CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector.
In *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 241–245.

[Pfeifenberger et al., 2019] Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2019).
Deep complex-valued neural beamformers.
pages 2902–2906.

Thank you for your attention

nicolas.furnon@loria.fr