GHENT UNIVERSITY

imec

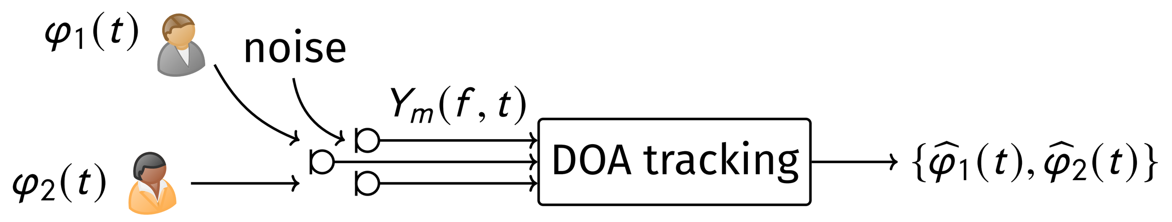
Improved Deep Speaker Localization and Tracking: Revised Training Paradigm and Controlled Latency

Alexander Bohlender, Liesbeth Roelens, and Nilesh Madhu

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium

Alexander.Bohlender@UGent.be, Liesbeth.Roelens@gmail.com, Nilesh.Madhu@UGent.be

1 Introduction



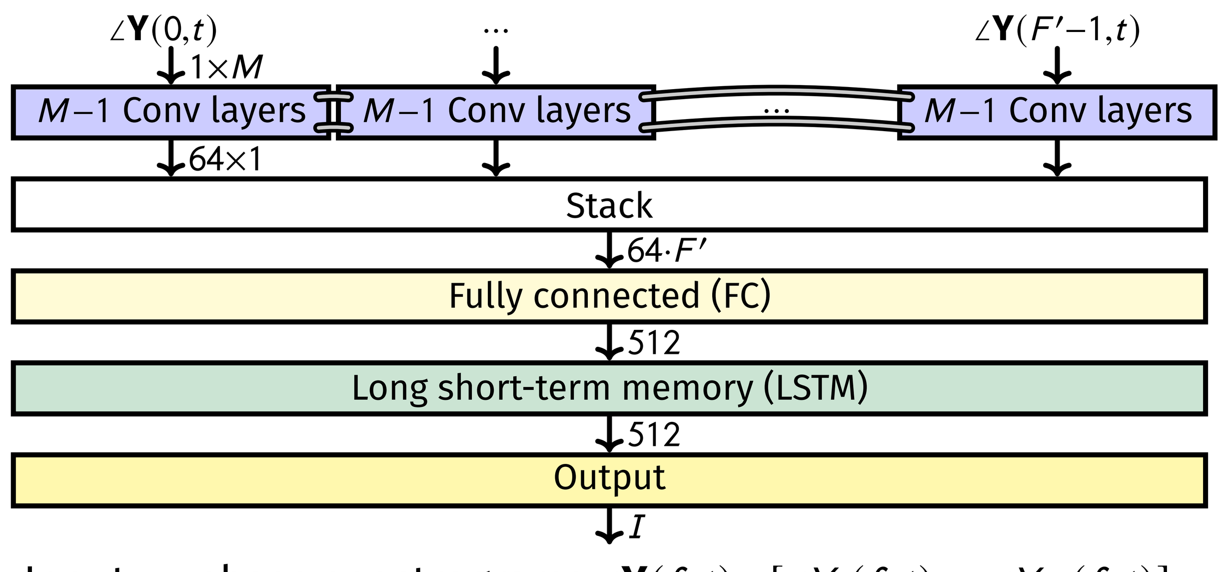
- Frequency index $f \in \{0, \dots, F'\}$, F': Nyquist
- Frame index t
- Source index $j \in \{1, \dots, J\}$ (above: $J = 2$)
- Microphone index $m \in \{1, \dots, M\}$
- (Azimuth) direction of arrival (DOA) $\varphi_j(t)$

Goal

Estimate and track the DOAs of *moving* talkers with a deep neural network (DNN) trained with simulated data.

2 Prior Work

Convolutional neural network (CNN) based on [1] with LSTM extension and training data generation of [2]



Input: phase spectrograms $\angle\mathbf{Y}(f,t) = [\angle Y_1(f,t), \dots, \angle Y_M(f,t)]$
→ DOA information in interchannel time differences

Output: posterior probabilities of source activity for each DOA of the discrete grid $\varphi \in \{0°, 5°, \dots, 355°\}$
→ classification problem with $I = 72$ classes

Training data generation

$$\mathbf{Y}(f,t) = \sum_{j=1}^{J} A_j(t)\,\mathbf{X}_j(f,t) + \mathbf{V}(f,t) \tag{1}$$

Activity $A_j(t)$: sources can be active ($A_j(t) = 1$) or inactive ($A_j(t) = 0$) at different times, transition between these two states with defined probability

Source $\mathbf{X}_j(f,t)$: time domain convolution of clean speech with simulated room room impulse responses (RIRs)

DOAs: newly selected every time a source becomes active ($A_j(t) = 1$, $A_j(t-1) = 0$)

Noise $\mathbf{V}(f,t)$: spatially diffuse but temporally uncorrelated, random source-to-noise ratio

Detecting sudden changes ✓

Modeled by source activity $A_j(t)$ and random DOA changes.

Tracking continuous trajectories of moving talkers ✗

Special case (jumps only between neighboring DOA classes), but not explicitly modeled.

3 Improved Moving Speaker Tracking

Simulation of moving speakers during training

Biased random walk model for jth source DOA:

$$\varphi_{j,q} = \varphi_{j,q-1} + D_{j,q}\,\Delta\widetilde{\varphi} \tag{2}$$

Segment q: fixed source location in each short segment $q \in \{1, \dots, Q_j\}$ → clean speech can still be convolved with pregenerated RIRs to obtain $\mathbf{X}_j(f,t)$ (no need for an online simulation of the room acoustics)

Direction $D_{j,q}$: movement in positive ($D_{j,q} = +1$) or in negative ($D_{j,q} = -1$) direction, direction changes with defined probability

Step size $\Delta\widetilde{\varphi}$: determined by the grid resolution (here 5°)

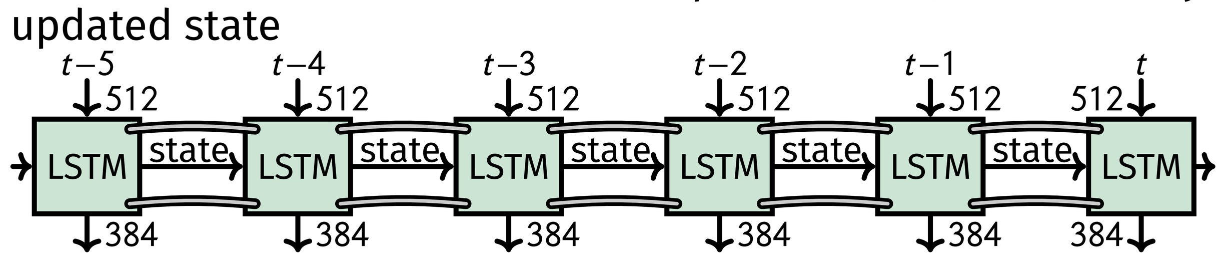
Motivation

Simple model permits easy online training data generation. Yet, accounting for different *angular velocities*, *source-array distances*, and *movement directions* still enables a good generalization to real-world scenarios.

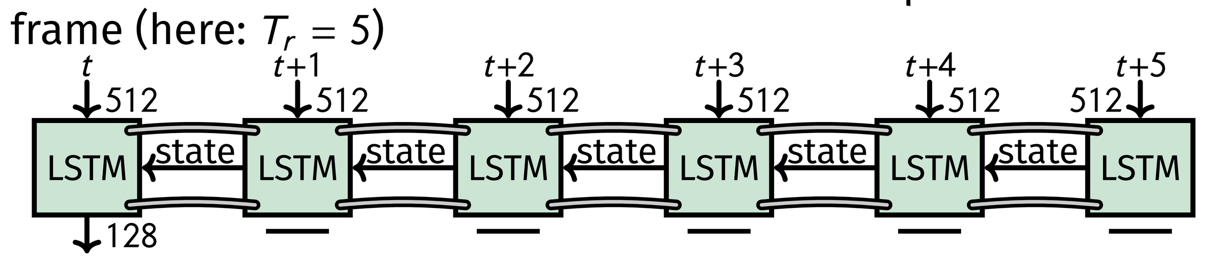
To cope with *both* types of DOA changes, embed gradual movements (2) into jumping sources framework (1).

Latency controlled bidirectional LSTM

Forward LSTM: unlimited context of past framers, continuously updated state



Backward LSTM: limited context of T_r future frames, state determined based on a different short subsequence in each frame (here: $T_r = 5$)



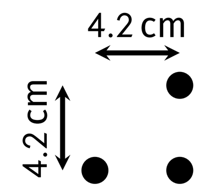
Preserve output dimensions: concatenate 384 features from forward, 128 features from backward ($\Sigma = 512$) → combination forms latency controlled bidirectional LSTM (LC-BLSTM) [3].

Motivation

DOAs typically change slowly over time → future context can be helpful. Controlled latency may still be acceptable for real-time applications.

4 Evaluation

$$\mathbf{Y}(f,t) = \sum_{j=1}^{J} \mathbf{X}_j(f,t) + \mathbf{V}(f,t)$$



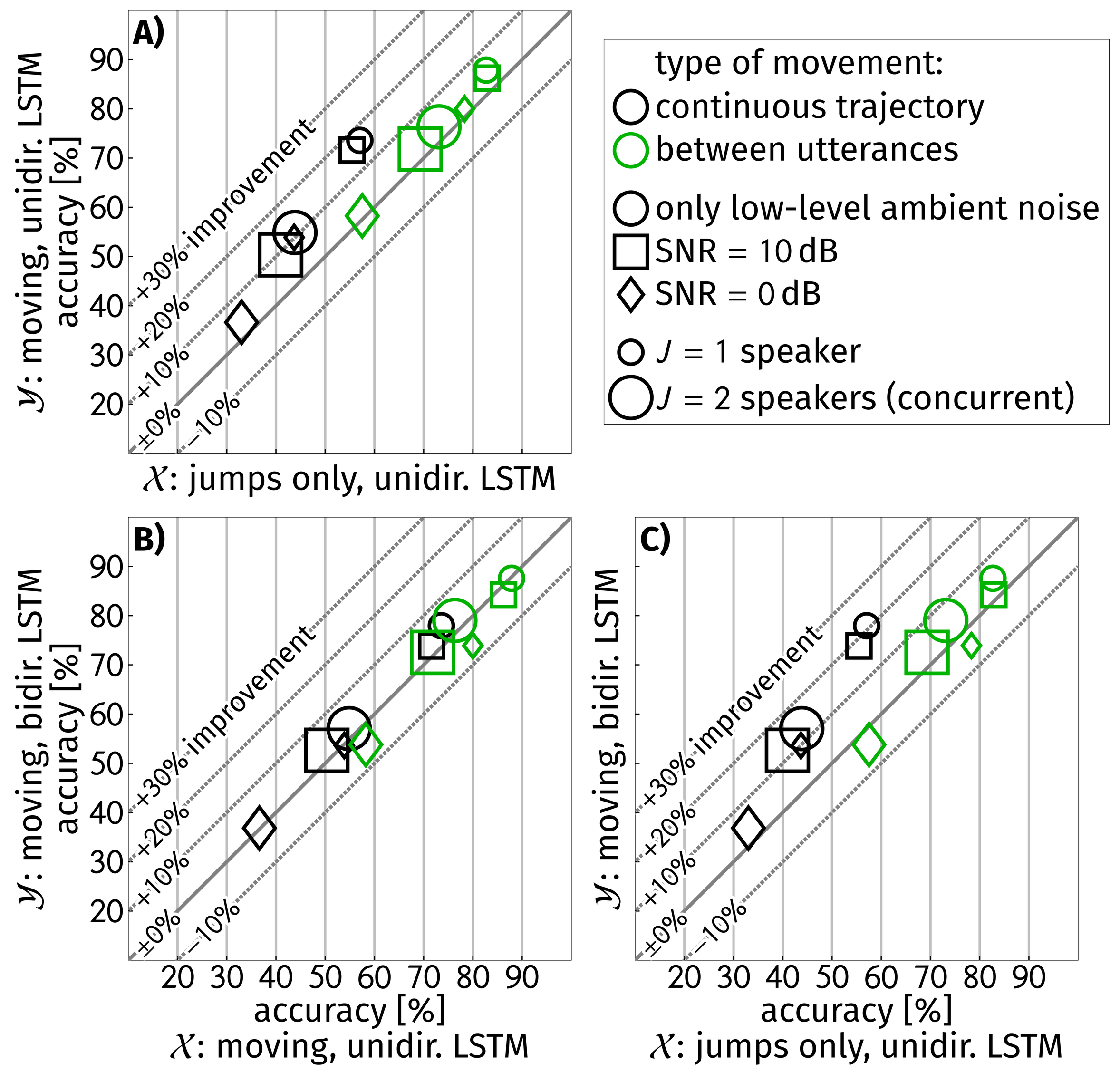
$\mathbf{X}_j(f,t)$: individually recorded 4 talkers in 3 rooms (12 in total) moving around a table in 2 different scenarios: 1) continuous movement while speaking, 2) walk several steps only between two utterances

$\mathbf{V}(f,t)$: relatively diffuse pub noise recording

array: triangular configuration of 3 microphones

The localization accuracy (fraction of correct DOA estimates) with a tolerated error of 7.5° is used to measure performance.

Quantitative analysis



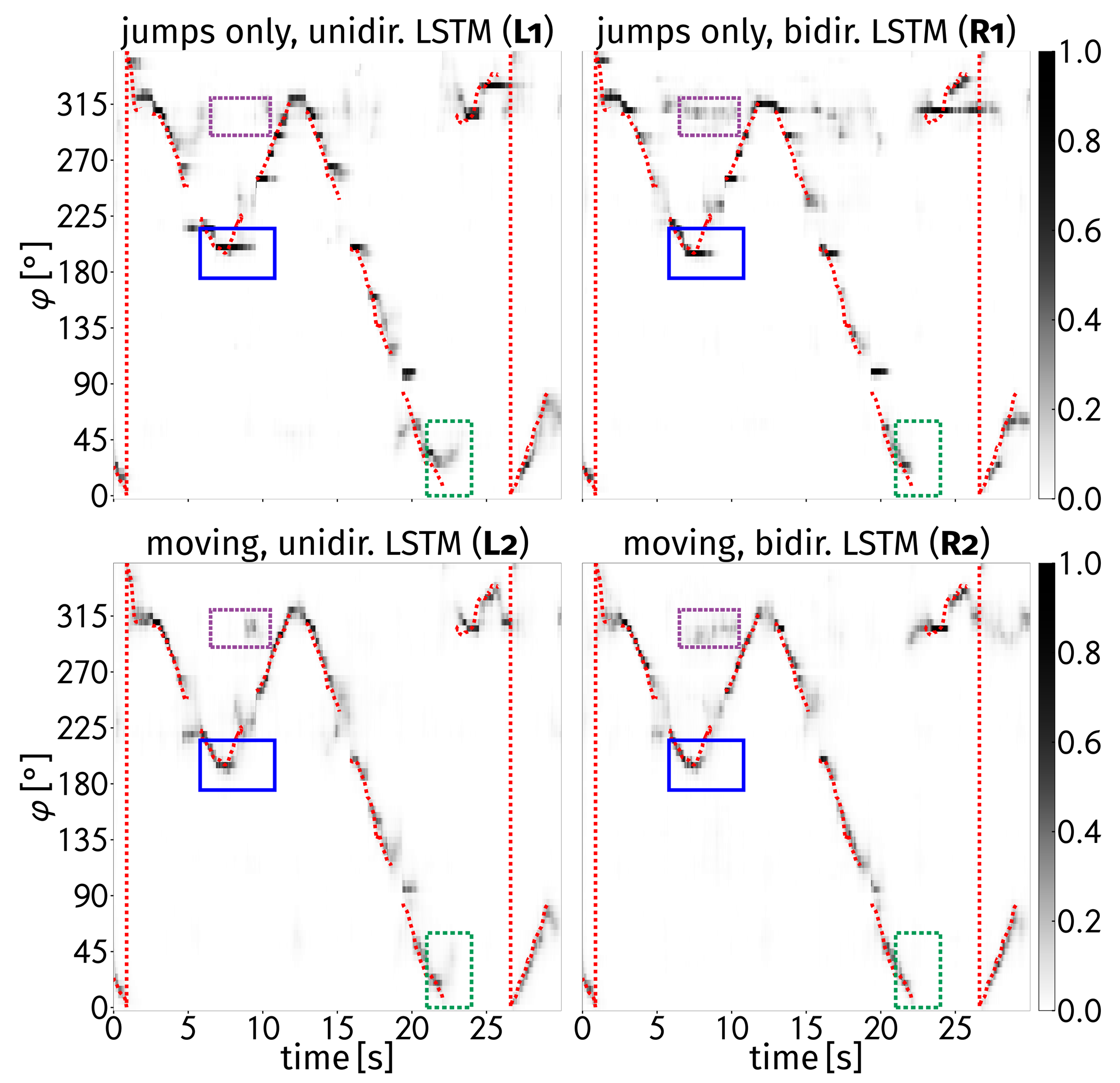
A) Training with gradual movements improves localization of moving talkers by 10-20%, scores do not deteriorate when talkers are stationary during speech activity

B) Minor but consistent increase (up to 5%) only of the *moving* talker localization accuracy

C) Further improvement by combining both modifications

Qualitative Analysis

DNN output (posterior probabilities) for an example with one talker and SNR = 5 dB (dotted red line is ground truth):



L1) Baseline: only updates estimate when true DOA has already changed significantly → delay of up to a few seconds in the tracking of a continuous movement

2 vs. 1) Smooth tracking of continuous source trajectories

R vs. L) Limited future context helps to, e. g., detect the end of an utterance more quickly, but adversely affects robustness (high probabilities at false locations)

5 Conclusions

- Model movement trajectories in training by small jumps between neighboring discrete DOAs
  - → Smooth tracking of real moving talkers
  - → Simple model is sufficient, no complex online simulation of the room acoustics is needed
- LC-BLSTM incorporates strictly limited future context
  - → Information from a small number of frames may be less reliable, could give rise to increased sensitivity to noise
  - → Moving talker localization still improves slightly overall

References

[1] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[2] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1594–1608, 2021.

[3] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5755–5759.