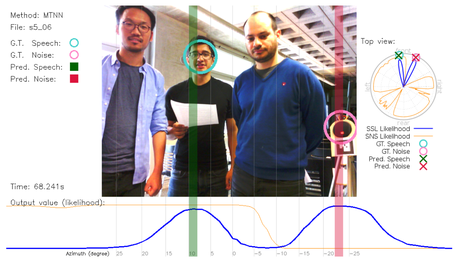
Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network

Weipeng He[1,2], Petr Motlicek[1] and Jean-Marc Odobez[1,2]

[1]Idiap Research Institute [2]École Polytechnique Fédérale de Lausanne (EPFL)



Problems: Localization and Classification

Localize sound sources and **classify** them into speech or non-speech sources in complicated human-robot interaction scenarios:

- Simultaneous sound sources
- Speech and non-speech sources
- Strong robot ego-noise

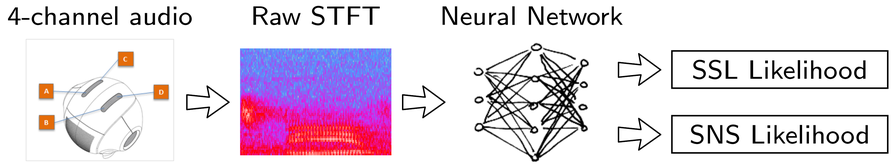
Contribution: Jointly Solving Both Tasks

Sound localization and classification can help each other:

- Localization provides spatial information for classification.
- Classification provides spectral information for localization.

However, previously the localization and classification are solved sequentially. We propose solving both tasks jointly using a multi-task neural network.

System Overview



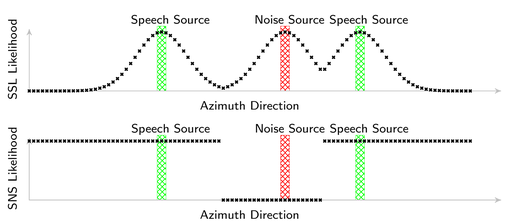
Output and Loss Function

The network outputs on each direction, the likelihood of the presence of a sound source (SSL likelihood, $\mathbf{p} = \{p_i\}$) and the likelihood of the sound being a speech source (SNS likelihood, $\mathbf{q} = \{q_i\}$).

Desired output:

SSL Likelihood Maximum of Gaussian functions centered at the DOAs of the ground truth sources.

SNS Likelihood 1 if the nearest source is speech.



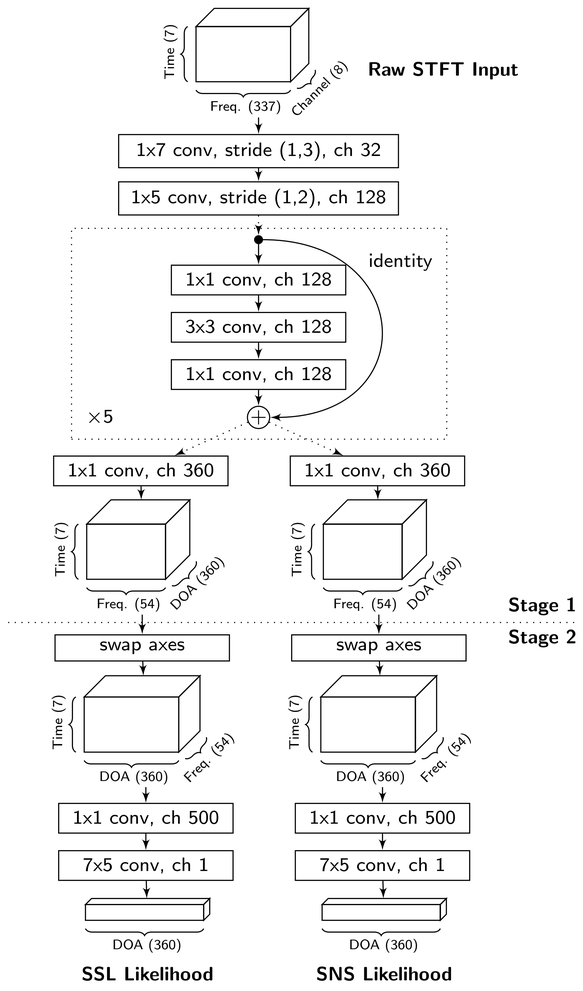
Loss function:

$$\text{Loss} = \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 + \mu \sum_i w_i \, |\hat{q}_i - q_i|^2$$

The SNS loss is weighted by $\{w_i\}$, which depends on its distance to the nearest source, so that the network is trained with the emphasis around the directions of the active sources.

Convolutional Neural Network

Fully convolutional neural network with residual network trunk and two task-specific branches:



Two-stage Training

Output of Stage 1 corresponds to local time-frequency area in the input because of the convolutions. So we train the network in two steps:

1. Supervision on Stage 1 to initiate early predictions on each TF point.
2. Train network end-to-end.

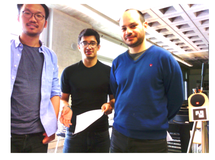
Data

Loudspeakers

- 32 hours train / 17 hours test
- Speech: AMI Corpus
- Non-speech: AudioSet

Human talkers

- 8 minutes test

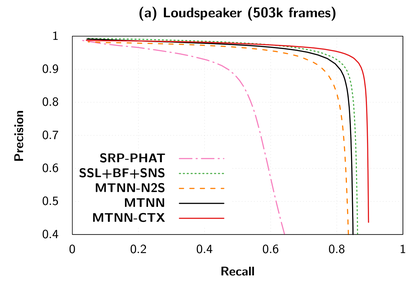


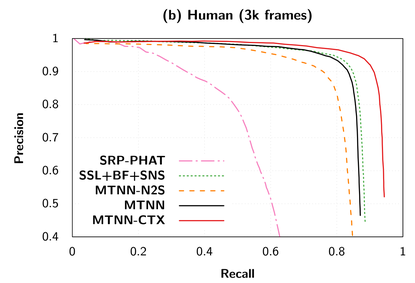
Methods

MTNN The proposed multi-task network.
MTNN-N2S The proposed methods without the two-stage scheme.
SSL+BF+SNS Sequentially localize, beamform and classify sounds with NN.
MTNN-CTX The proposed methods with temporal context.

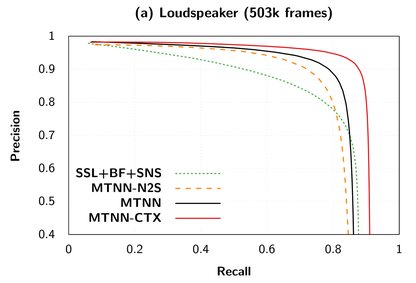
Results

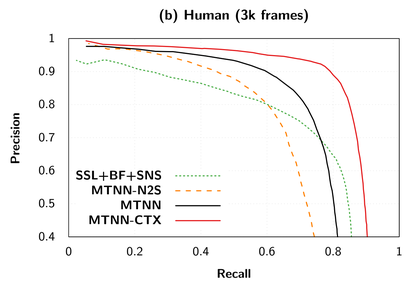
Sound Localization





Speech Localization





Speech/Non-speech Classification

Dataset	Loudspeaker	Human
SSL+BF+SNS	0.80	0.68
MTNN-N2S	0.93	0.82
MTNN	**0.95**	**0.85**
MTNN-CTX	0.96	0.89

Conclusion

- Significant better performance compared to SSL+BF+SNS in classification and speech localization.
- Further improvement by adding temporal context.