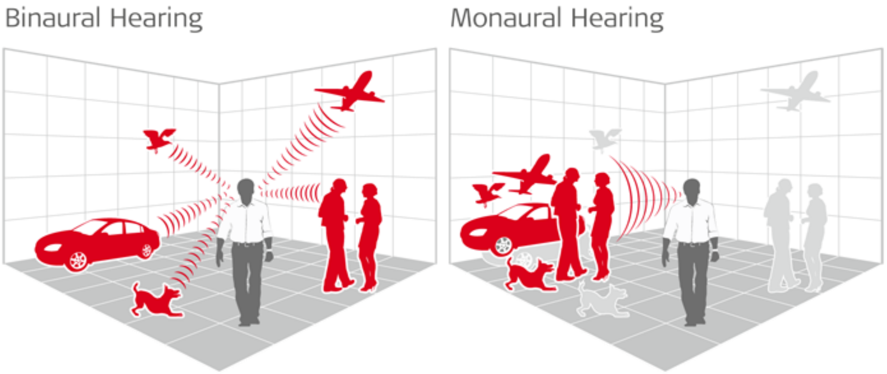
ECE 477: Sound Source Localization

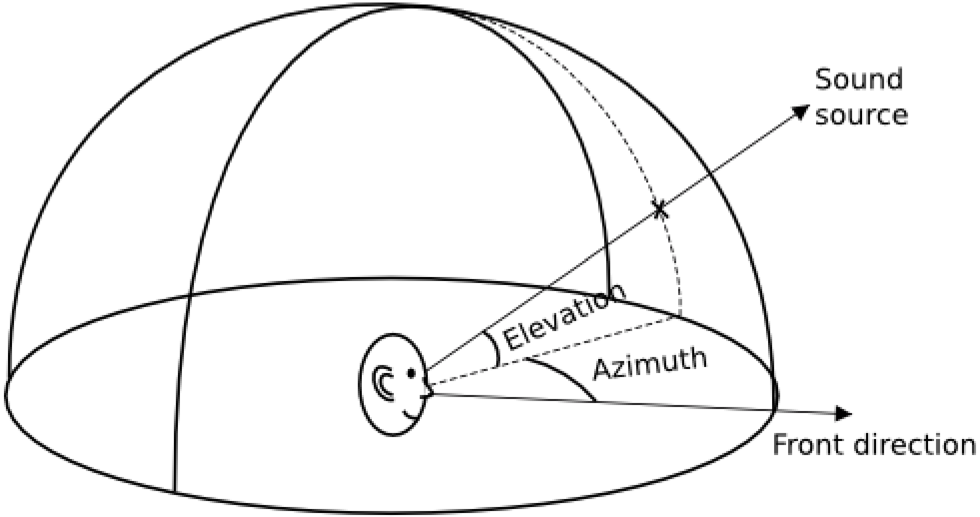
Shafaqat Rahman and John Kyle Cooper



Outline:

1) Sound Localization and Its Importance
2) Early History of Sound Localization
3) Current Approaches and Algorithms
    a) Cross-Channel Algorithm
    b) Monaural Source Separation for Localization
    c) Deep Convolutional Neural Networks (DCNN)
    d) Probabilistic Neural Networks (PNN) and Cross-Correlation
    e) Comparison of Main Approaches
4) Limitations of Past and Present
5) Future Directions

Aim of Sound Source Localization



Human Auditory Model → Computer Auditory Scene Analysis

IMPORTANCE

Difficulties Experienced By Hearing Aid and Cochlear Implant Users

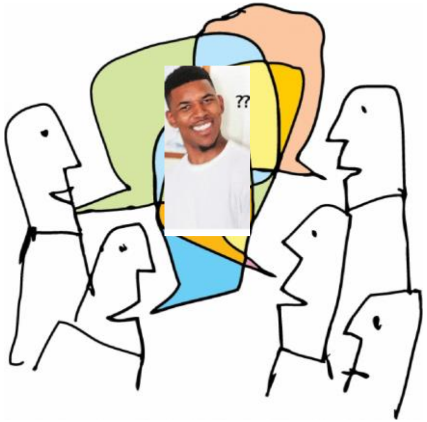
Speech Understanding in Noise

Localization in Noise

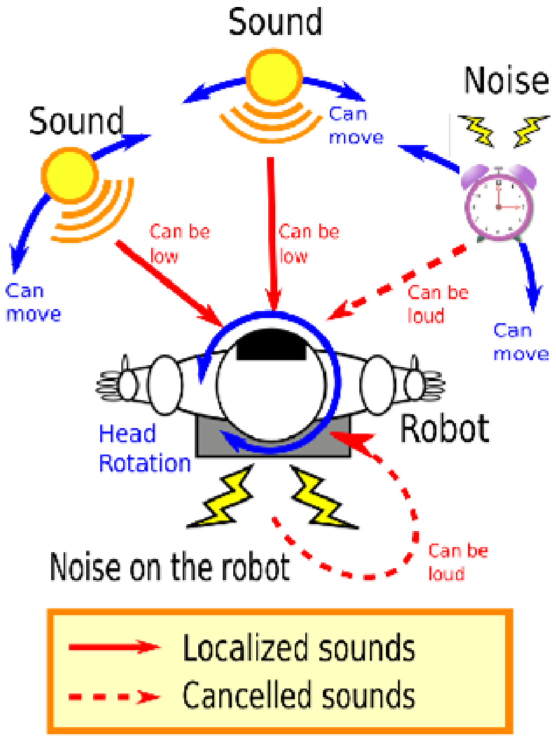
Sound Source Localization in Robotics

Human Interaction

Cochlear Implant User



“Cocktail Party Problem”



1. Kerber and Seeber (2012)
2. Rascon et al. (2017)

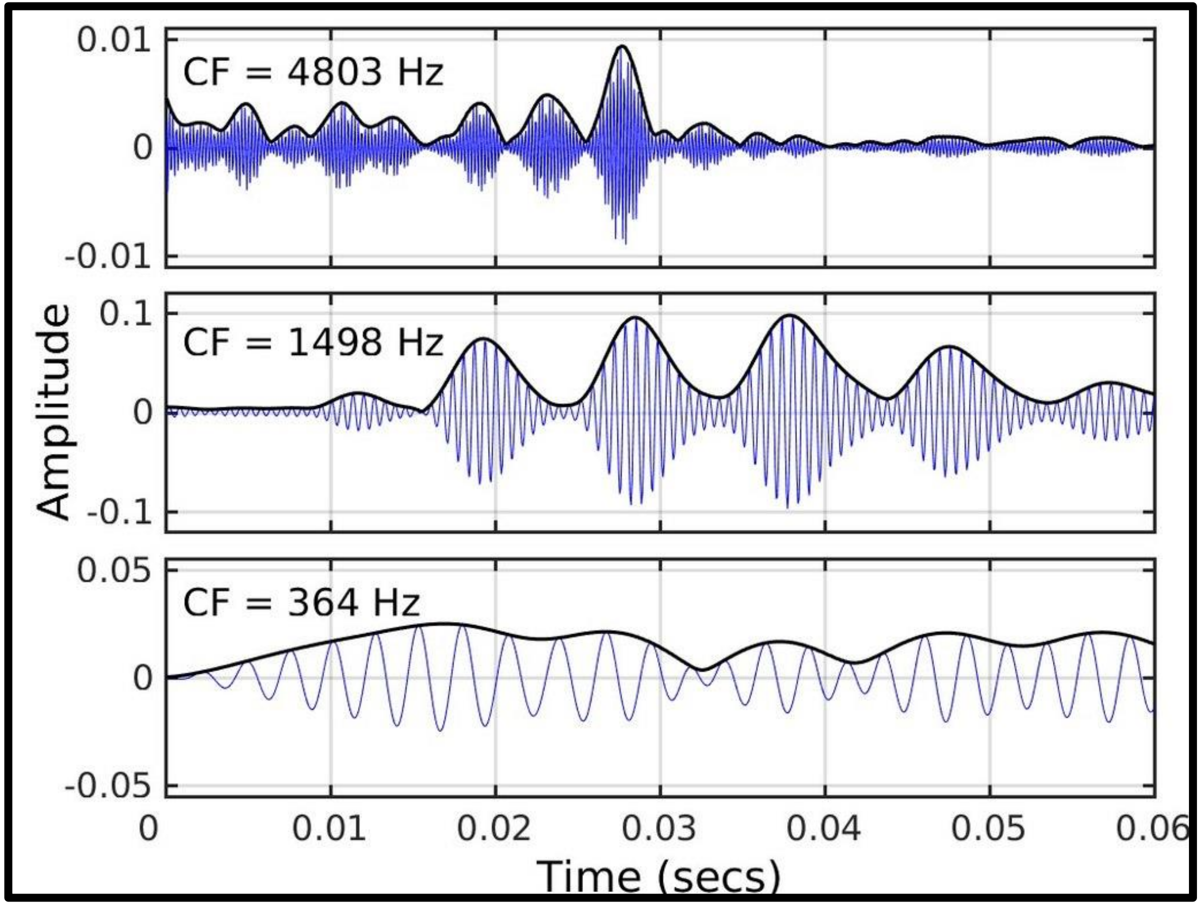
Acoustic Cues

Interaural Level Differences (ILDs)

Interaural Time Differences (ITDs)

in the envelope (speech)

in the waveform (pitch & music)



1. Kerber and Seeber (2012)
2. Michael Stone (2018)

Early History of Sound Localization:

XII. On Our Perception of Sound Direction - Lord Rayleigh, 1907
The localization of actual sources of sound - S.S Stevens and E.B. Newman, 1934

Early papers establishing the basis behind ITD and ILD

Low frequency sounds are accounted by phase differences (ITD)

High frequency sounds are accounted for by level differences (ILD)

Intermediate range where localization is poorest

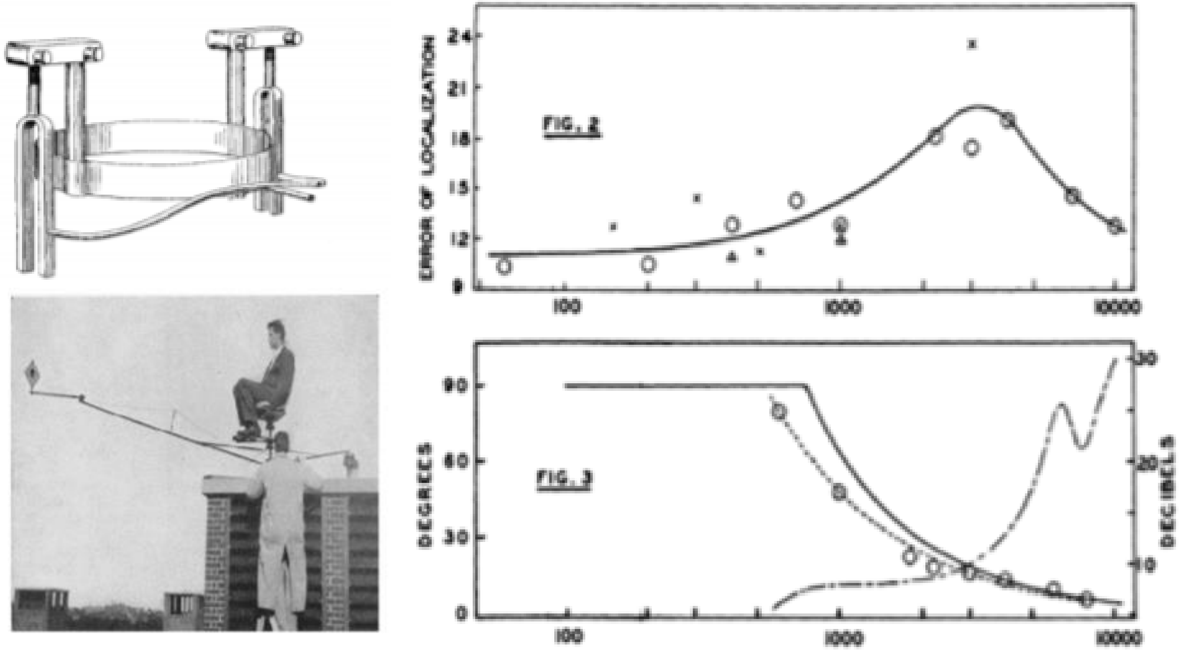


Figure 1: A) Early designed tuning forks attached to the head for localization studies **B**) stimulus generation of pure tones conducted on a pedestal (on the roof of the Harvard Biology building) **C)** Plots depicting error of localization in respect to phase and level.

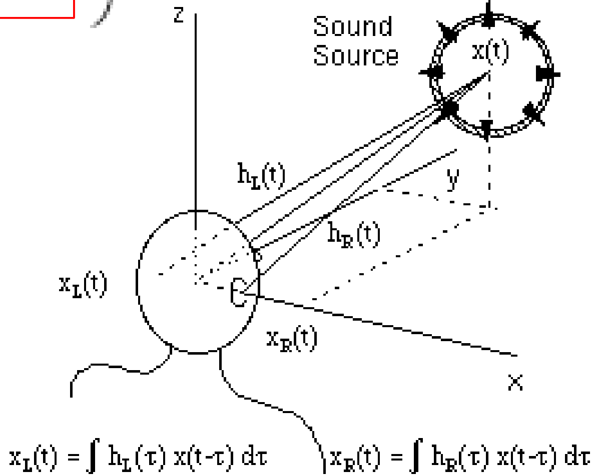
MacDonald & Tran (2006): Cross-Channel Sound Localizer

Head-Related Transfer Functions (1 for each microphone,).

Inverse algorithm

$$\min_{\hat{\theta},\hat{\phi}} \sum \left(\boxed{R_{Left}} * \boxed{\left[H_{Left}^{(\hat{\theta},\hat{\phi})}\right]^{-1}} - \boxed{R_{Right}} * \boxed{\left[H_{Right}^{(\hat{\theta},\hat{\phi})}\right]^{-1}} \right)^2$$

Digital Recordings of sound from left and right microphones (in ear canal of mannequin)



MacDonald & Tran (2006): Cross-Channel Sound Localizer

Head-Related Transfer Functions (1 for each microphone,).

Using commutability and transitivity of convolution operator...

$$\min_{\hat{\theta},\hat{\phi}} \sum \left(\boxed{R_{Left}} * \boxed{H_{Right}^{(\hat{\theta},\hat{\phi})}} - \boxed{R_{Right}} * \boxed{H_{Left}^{(\hat{\theta},\hat{\phi})}} \right)^2$$

Essentially "convolves" each recording by the transfer function of the opposite microphone and obtains location coordinates at minimum

Digital Recordings of sound from left and right microphones (in ear canal of mannequin)

MacDonald & Tran (2006): Cross-Channel Sound Localizer

Strengths

Highly accurate in quiet environments with 2-sensors and works well in mildly noisy environments with 4-sensors

Frequency based cues minimize error caused by front-back reversals

Issues

Head Related Transfer Functions are predetermined, but that's not the case in practice

Performance is likely to decrease in reverberant environments

Experiments were done at fixed sound source elevation. Unsure of how results will be when elevation is varied.

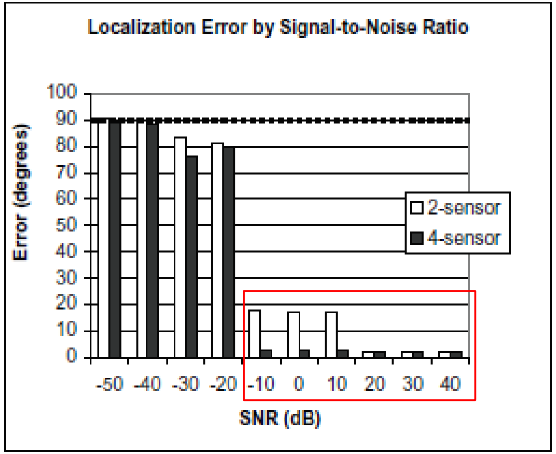


Figure 1

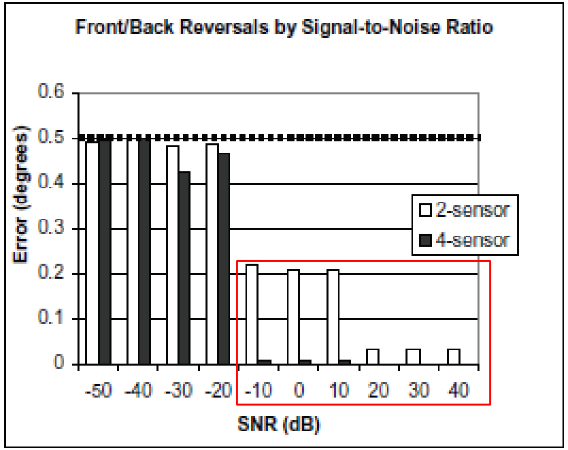


Figure 2

Carlos et al. (2013):
Spherical Cross-Channel Algorithm for Binaural Sound Localization

Assuming the HRTFs of the robot's head are known is apparently "impractical", so a spherical generalization is made.

FT of a source signal emitted from a punctual source

Head-related transfer functions positioned at a specific point

$$I_L(r,\theta,\varphi,\omega) = \boxed{H_L(r_s,\theta_s,\varphi_s,\omega)}\boxed{S(\omega)}\boxed{H_R(r,\theta,\varphi,\omega)}$$
$$I_R(r,\theta,\varphi,\omega) = H_R(r_s,\theta_s,\varphi_s,\omega)S(\omega)H_L(r,\theta,\varphi,\omega)$$

Source Position Estimation

$$(\hat{r}_s,\hat{\theta}_s,\hat{\varphi}_s) = \arg\max_{r,\theta,\varphi}\left\{\boxed{\mathrm{corr}\left[(I_L(r,\theta,\varphi,\omega), I_R(r,\theta,\varphi,\omega)\right]}\right\}$$

Traditional Pearson's correlation coefficient

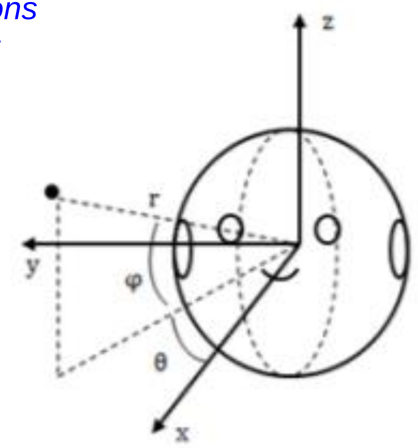


Fig. 1. Front view of the spherical head. A point's position is defined according to the standard LISTEN spherical coordinates system with the distance r, the azimuth θ and elevation φ [13].

Carlos et al. (2013):
Spherical Cross-Channel Algorithm for Binaural Sound Localization

Assuming the HRTFs of the robot's head are known is apparently "impractical", so a spherical generalization is made.

FT of a source signal emitted from a punctual source

Head-related transfer functions positioned at a specific point

$$I_L(r,\theta,\varphi,\omega) = \boxed{H_L(r_s,\theta_s,\varphi_s,\omega)}\boxed{S(\omega)}\boxed{H_R(r,\theta,\varphi,\omega)}$$

$$I_R(r,\theta,\varphi,\omega) = H_R(r_s,\theta_s,\varphi_s,\omega)S(\omega)H_L(r,\theta,\varphi,\omega)$$

$$H^s(r,\beta,\omega) = \frac{P_s(r,\beta,\omega)}{P_f(r,\omega)} = \frac{rce^{-jr\omega/c}}{ja^2\omega}\sum_{m=0}^{\infty}(2m+1)P_m\left[\cos(\beta)\right]\frac{h_m(r\omega/c)}{h'_m(a\omega/c)}$$

$$H_L^s(r,\theta,\omega) = H^s\left(r,-\frac{\pi}{2}-\theta,\omega\right),$$

$$H_R^s(r,\theta,\omega) = H^s\left(r,\frac{\pi}{2}-\theta,\omega\right).$$

c = speed of sound
a = head radius
P_m =legendre polynomial
H_m = Hankel functions

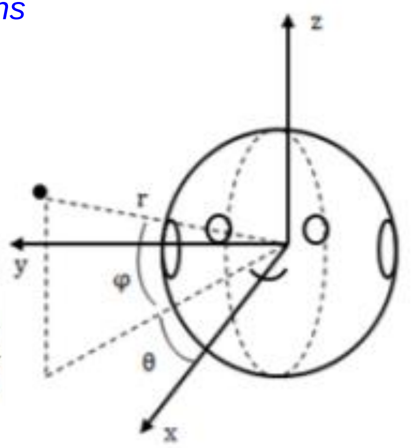


Fig. 1. Front view of the spherical head. A point's position is defined according to the standard LISTEN spherical coordinates system with the distance r, the azimuth θ and elevation φ [13].

Carlos et al. (2013):
Spherical Cross-Channel Algorithm for Binaural Sound Localization

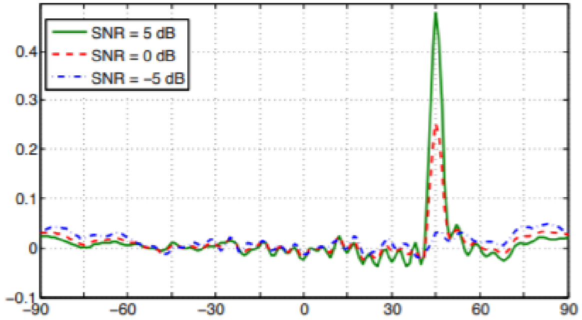


Fig. 3. Correlation coefficient as a function of the azimuth θ for $\mathrm{SNR_{dB}} = \{-5, 0, 5\}$dB, with a source emitting from $\theta_s = 45°$ and a spherical head.

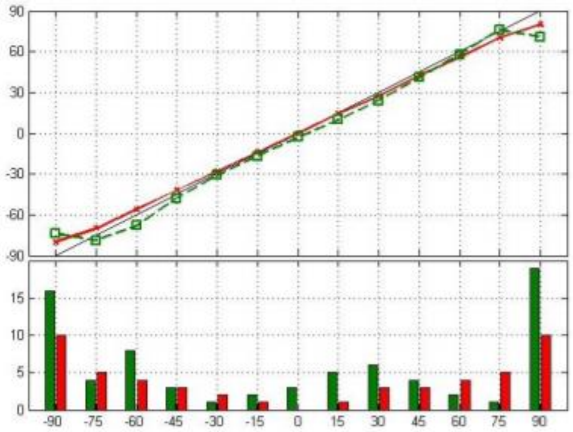


Fig. 7. (Top) Experimental mean estimated azimuth $\hat{\theta}$ as a function of the real angle for a spherical head (red) or a KEMAR head (dotted, green). (Bottom) Absolute value of the mean angular error for the two heads.

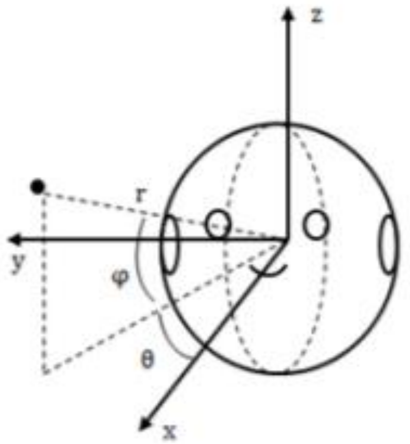


Fig. 1. Front view of the spherical head. A point's position is defined according to the standard LISTEN spherical coordinates system with the distance r, the azimuth θ and elevation φ [13].

Spherical approximation of the head works very well at the 13 tested positions with a 4º mean angular error (error at 10º in lateral directions).

Spherical accuracy > KEMAR head accuracy

Woodruff et al. (2012): Monoaural Source Segregation

Binaural Pathway

Azimuth-dependent Gaussian Mixture Model of interaural time differences (ITDs) and interaural level differences (ILDs).

Monaural Pathway

Multipitch Tracking (HMM-based)

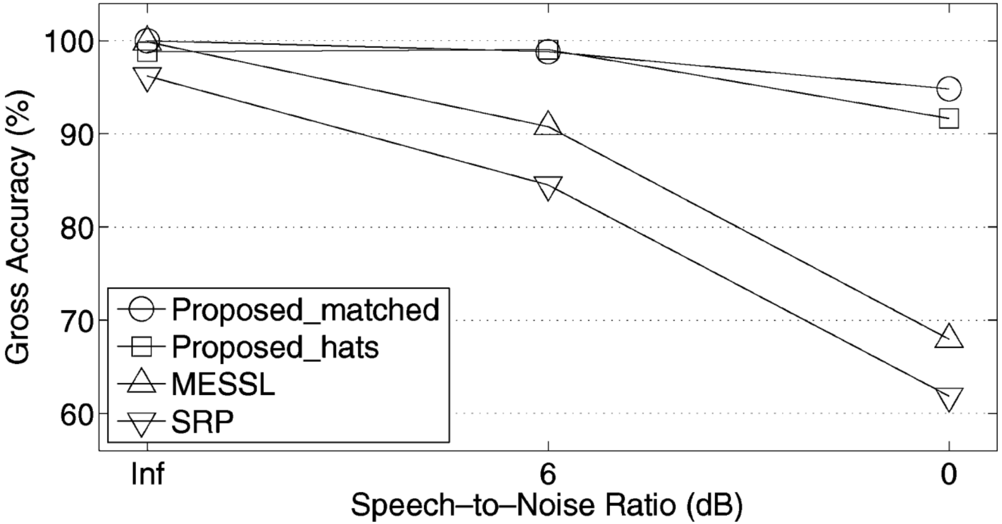
Pitch-Based Grouping

Onset/Offset Based Segmentation

Localization Framework

Maximum likelihood

Head and Torso Simulator (HATS)



Gross accuracy (%) as a function of noise level for the HATS evaluation set.

Yalta et al. (2016):
Sound Source Localization Using Deep Learning Models

Input

Data Preparation: Audio stimuli consisted of Japanese utterances (216 training, 72 evaluation). Training was done on HEARBO robot

STFT(N-channel audio → Power info was extracted and normalized at each frequency bin (W)

STFT frames were stacked to dimension H. Final input to DCNN is a N x W x H dimensional file.

Output

STFT of for one channel is obtained to evaluate its RMS power. If frames have RMS = -120, it's labeled "silent". Otherwise, the target label is set to the detected angle of the receiving audio

Calculated Correct Detection Rate and Correct Accuracy Rate for Evaluation

Yalta et al. (2016): Sound Source Localization Using Deep Learning Models

Training and Preparation (Continued)

- **HEARBO robot was equipped with microphone array (8 and 16 channels)**

- **Experiment setting: 4x7 m room with 200 ms of reverberation every 5 degrees.**

- **33750 files were selected for training at EACH angle (total preparation of audio files for 360 degrees = 2,430,000).**

- **Distribution include sound mixtures with and without noise (Clean, 30, 10, 5, 0, -5, -10, -20, and -30 dB).**

Loss function: Cross Entropy

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

M = number of classes (in this case, 360 angles)
Y = binary indicator (0 or 1) dependent on if angle 'c' is correct to observation 'o'
P = predicted probability

Yalta et al. (2016):
Sound Source Localization Using Deep Learning Models

Residual learning speeds up the learning process and improves training convergence.

Made variations of the general DCNN architecture (Plain vs.. ResNet1..ResNet4). Compared to SEVD-MUSIC

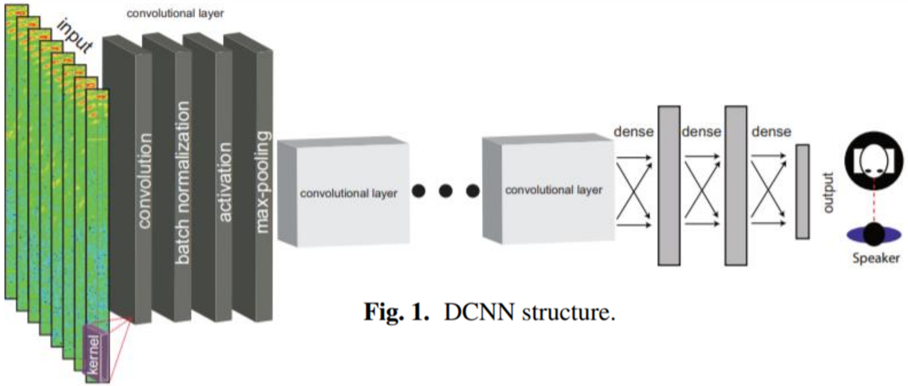


Fig. 1. DCNN structure.

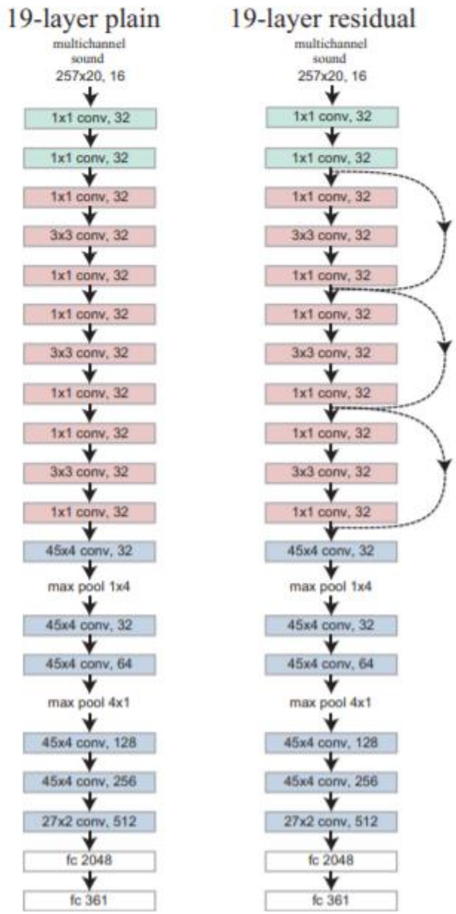


Fig. 4. Network architecture. Left: plain network. Right: residual network. The dotted line shortcut represents a 1×1 convolutional layer.

Yalta et al. (2016): Sound Source Localization Using Deep Learning Models

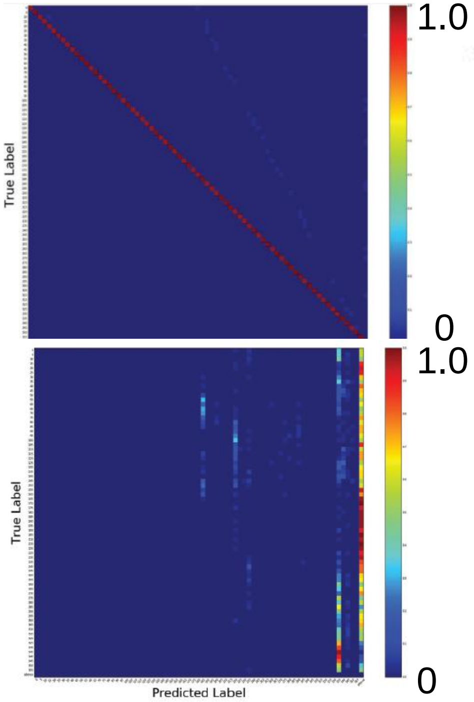


Fig. 9. ResNet4 angle prediction confusion matrix. Top: Clean. Bottom: −35 dB SNR.

Table 4. Microcone (array of 7 microphones) block accuracy.

SNR (dB)	SEVD-MUSIC	ResNet1
−35	**1.39**	1.39
−30	**1.39**	1.39
−25	**1.39**	1.39
−20	**1.42**	1.39
−15	**2.64**	2.06
−10	**9.10**	6.42
−5	**21.16**	16.94
0	**36.12**	33.61
5	**50.99**	49.69
10	**66.75**	64.61
15	**81.26**	78.11
30	**97.42**	95.36
45	**98.01**	97.39
Clean	**98.17**	97.50

Table 6. HEARBO accuracy rate.

SNR (dB)	PlainNet	ResNet1	ResNet2 TanH/ELU	ResNet3	ResNet4
−35	−99.40	−0.05	−21.47	−17.24	−60.15
−30	−93.91	2.83	−16.69	−16.03	−32.75
−25	−84.83	7.97	−4.35	−12.15	−12.46
−20	−61.85	19.46	15.38	−4.30	1.08
−15	−13.04	40.05	36.14	−2.69	14.3
−10	21.18	55.81	46.15	23.60	14.58
−5	38.68	61.73	50.88	41.70	11.98
0	48.43	62.96	53.53	49.75	18.81
5	54.51	63.07	56.80	55.80	33.72
10	58.74	63.51	59.45	60.42	48.63
15	62.40	65.50	61.94	63.77	59.06
30	68.06	69.60	69.24	70.08	69.82
45	71.04	71.70	72.30	72.04	72.11
Clean	70.89	71.53	72.01	71.90	71.46

Confusion matrix y-axis = probability from 0 to 1.0

Sun et al. (2018): Generalized Cross Correlation Classification Algorithm (GCA)

Nonlinear Classification Problem

$X = H \otimes s(t) + n_m(t)$

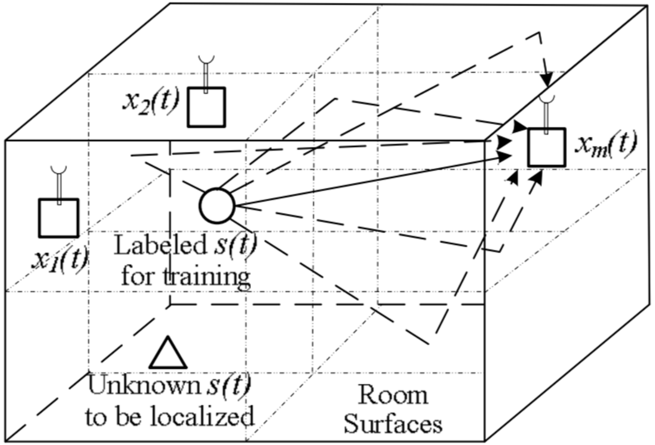
$c_s = \text{classify}(X, \sum \text{feature}_i)$

Probabilistic Neural Network (PNN)

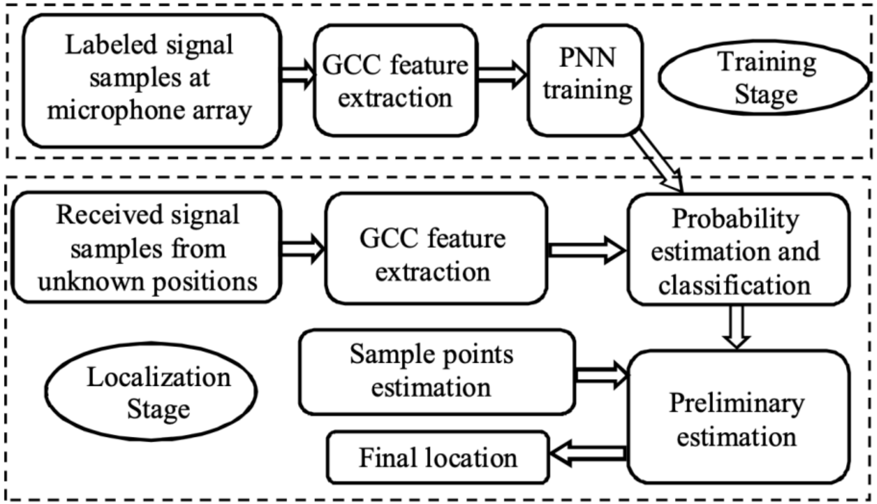
- input layer
- pattern layer
- summation layer
- decision layer

Generalized Cross Correlation (GCC)

At the beginning of the PNN training, the enclosed room is divided into a number of *K* equal-dimension rectangular clusters.



Space Cluster Classification for SSL



Flow Chart of The Proposed GCA

Comparison of Main Approaches (Methodology)

MacDonald & Tran (2006): Cross-Channel Algorithm

Pros: Models auditory system of humans by using acoustic cues (ILDs and ITDs) and frequency related cues to estimate SSL via a control system's approach
Cons: Measures HRTF per estimation, which is impractical in real world applications

Woodruff et al. (2012): Monaural Source Separation

Pros: Uses monaural multipitch tracking to distinguish multiple sound sources in order to assist a binaural input with azimuth estimation.

Carlos et al. (2013): Spherical Cross Correlation

Pros: Relies on spherical generalization for modeling HRTFs. **Cons: (1)** spherical modeling cannot account for front-back reversals. **(2)** Humans don't have spherical heads.

Yalta et al. (2016): DCNN

Pros: Uses DCNN with real world impulse response and end-to-end training to perform sound source localization.
Con: Current architecture is still not optimized, needs to be further explored. Current training might now allow for angle detection below 1º

Sun et al. (2018): GCA

Pros: Extracts generalized cross correlation (GCC) features and then estimates sound source location using probabilistic neural network (PNN).

Comparison of Main Approaches (Big Picture Results)

MacDonald & Tran (2006): Cross-Channel Algorithm

Algorithm performs well above chance levels up to a SNR of -10 dB.

Woodruff et al. (2012): Monaural Source Separation

Algorithm is able to localize multiple sound sources well above chance level up to a SNR of 0 dB.

Carlos et al. (2013): Spherical Cross Correlation

Algorithm can localize sounds in 3 dimensions well above chance level up to a SNR of 0 dB.

Yalta et al. (2016): DCNN

DCNN model with residual layers (ResNet1, ResNet2) performs well above chance level up to a SNR -5 dB.

Sun et al. (2018): GCA

For 20 different acoustic environments, GCA can localize sounds in 3 dimensions well above chance level up to SNR of -10 dB and a T_{60} of 600 ms.

Limitations of Past and Current Approaches

Past Approaches

Past approaches were only able to localize the azimuth angle of the sound source

Current Approaches

Current approaches must balance the model's accuracy and computational complexity.

Current approaches cannot perform accurately above an SNR of -10 dB and do not perform well with reverberation times of 400 and 600 ms.

Future Research Directions

Future Improvements

Dynamic Acoustic Environment
Few studies are working on this challenge.

Few datasets devoted to sound source localization.
More training datasets will improve performance of future neural network models.

Future Applications

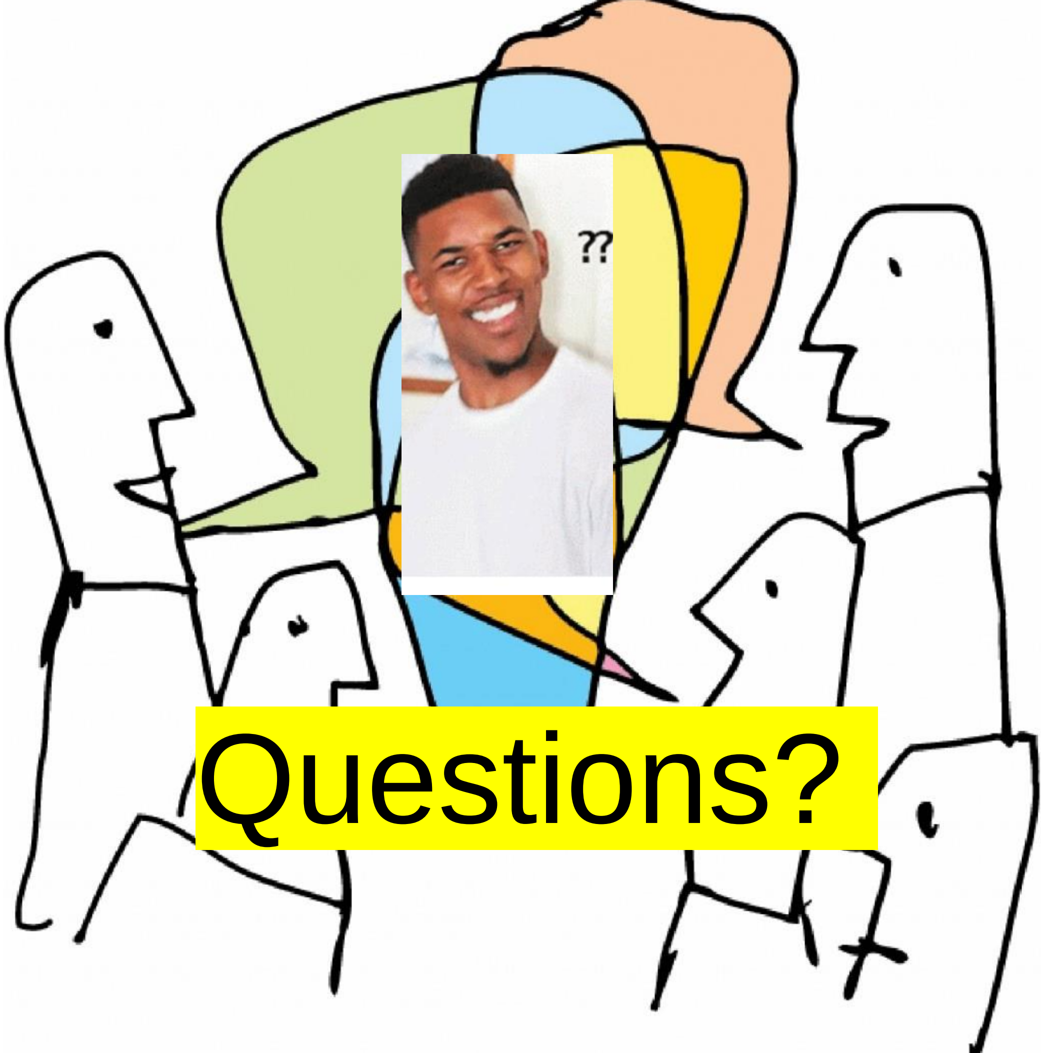
Simple Integration into existing platforms

SSL implementation in different types of robots

Integration of SSL in robotic tasks

Service-Robots for Elderly Care

1. Rascon et al. (2017)



Questions?

References

1. Rascon, C., & Meza, I. (2017). Localization of sound sources in robotics: A review. Robotics and Autonomous Systems, 96, 184-210. doi:https://doi.org/10.1016/j.robot.2017.07.011
2. Kerber, S., & Seeber, B. U. (2012). Sound localization in noise by normal-hearing listeners and cochlear implant users. *Ear Hear, 33*(4), 445-457. doi:10.1097/AUD.0b013e318257607b
3. Rayleigh, L. (1907). XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 13*(74), 214-232. doi:10.1080/14786440709463595
4. Stevens, S. S., & Newman, E. B. (1936). The Localization of Actual Sources of Sound. *The American Journal of Psychology, 48*(2), 297-306. doi:10.2307/1415748
5. MacDonald, J., & Tran, P. (2006). A Sound Localization Algorithm for Use in Unmanned Vehicles.
6. Vina, C., Argentieri, S., & Rébillat, M. (2013). *A Spherical Cross-Channel Algorithm for Binaural Sound Localization*.
7. J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503-1512, July 2012. doi: 10.1109/TASL.2012.2183869
8. Yalta, N., Nakadai, K., & Ogata, T. (2017). Sound Source Localization Using Deep Learning Models. *Journal of Robotics and Mechatronics, 29*, 37-48. doi:10.20965/jrm.2017.p0037
9. Y. Sun, J. Chen, C. Yuen and S. Rahardja, "Indoor Sound Source Localization With Probabilistic Neural Network," in IEEE Transactions on Industrial Electronics, vol. 65, no. 8, pp. 6403-6413, Aug. 2018. doi: 10.1109/TIE.2017.2786219
10. Zhu, N., & Reza, T. (2019). A modified cross-correlation algorithm to achieve the time difference of arrival in sound source localization. *Measurement and Control*, 002029401982797. doi:10.1177/0020294019827977