Swapnil Das, Soham Deshmukh, Abhishek Das, Revanth Banala Nithyanandam
Department of ECE, Carnegie Mellon University

Carnegie Mellon University
Electrical & Computer Engineering

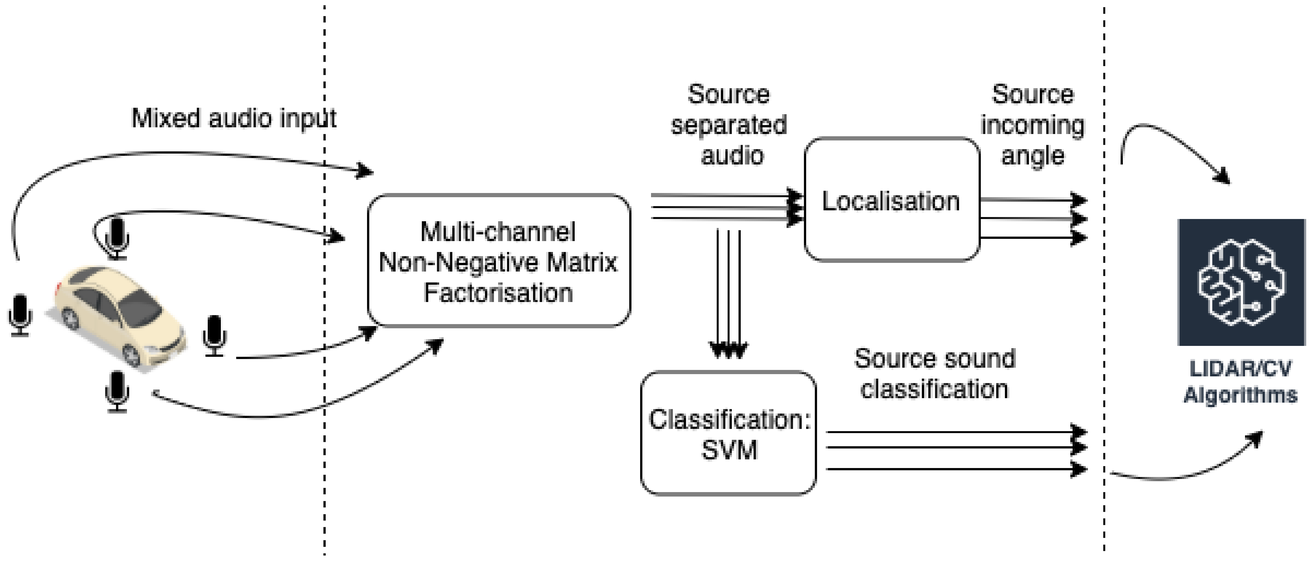
"Giving Cars Ears"

Acoustic Localization and Classification of Sound Sources for Autonomous Vehicles

Motivation

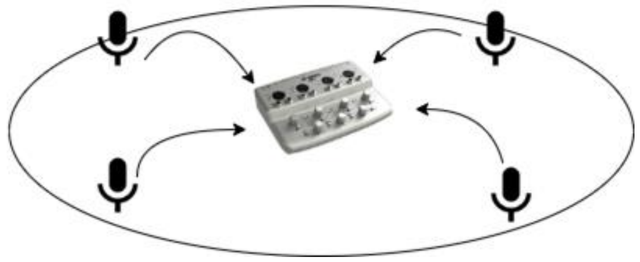
Audio-based perception for autonomous cars is very important for discerning and comprehending the environment, as it can **complement existing vision-based perception** and pave the way for multimodal operation. However, these sensors have not yet been fully established in modern autonomous vehicles, robbing the system of context when needing to make important driving decisions.

Our project seeks to take steps towards tackling this by **applying machine learning techniques** to help **separate, classify and localise** sounds of objects in a car's vicinity. With these two capabilities, a car is more well-informed of its surroundings and can act accordingly (such as responding and **making way for police cars, ambulances and fire trucks**).



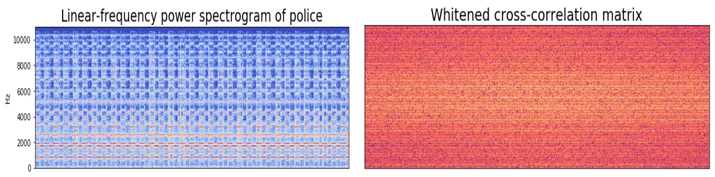
In our system, mixed audio inputs are separated via Multi-channel NMF. Separated sounds are classified as one of 'sirens', 'children', 'dogs' using 34 extracted features. We expect real-world applications in autonomous vehicles will use this data to corroborate vision perception.

System Setup



We captured microphone inputs with an **Alesis IO4 with four channels**. We need four channels because sound is read differently by each mic which is used for localisation; **each mic encodes unique spatial and temporal information**. For sound classification, we trained on the UrbanSound dataset.

Acoustic Localisation of Sound Sources

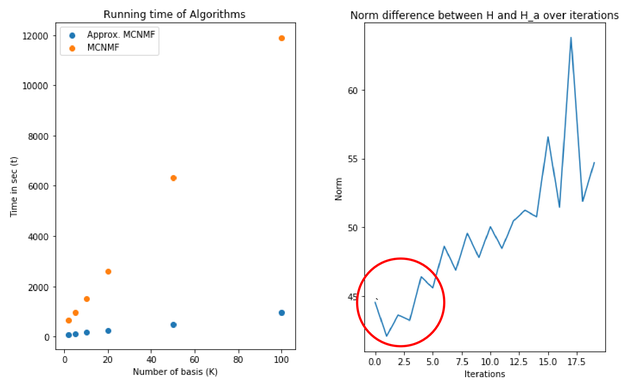


Whitened Cross-correlation spectrogram and PSD of the the acoustic signals from two mics (Sound: Police Siren, Correlation between phase variant signals from microphones 1 & 2)

Phase variant acoustic signals are modelled to estimate whitened **cross correlation** between signals with **spectral weights**. Using the distances between mics and **Time-Difference-of-Arrival** (TDOA) in 3D space, **locations of sound sources** are quantified **relative to the car**.[1]

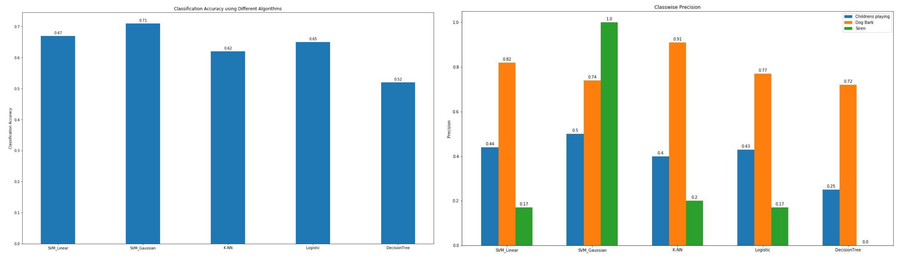
Multi-channel NMF

Multi-channel NMF[2] (MNMF) varies from the traditional algorithm as it **encodes signal magnitude** and **phase differences** between channels in a matrix H (which we apply in localisation). However, MNMF has a **long running time**, which deters real-time usage. To combat this, we developed an **MNMF approximation**, where all channel inputs are stacked with respect to time so that **bases and weights are learnt with traditional NMF**. The learnt data is **fed back into MNMF** to learn an approximated H we call H_a.



We see how we get the **best approximations with few iterations** of MNMF (as highlighted by the red circle in the plot to the right). Since the bulk of computation is in calculating H each iteration with a Riccati solver, we **achieve high speedup** with a smaller iteration count than pure MNMF (as shown in the plot to the left) and obtain a reasonable H_a to use for localisation.

Classification Results and Future Work



- Because of the classes we chose, the number of training samples we had was limited. **Collecting more data** could help **improve the accuracies** significantly.
- We could also **add more classes** to boost functionality - some key urban elements to look out for are trains and horns.
- In future work, we can **incorporate visual feedback as a prior** (e.g. through LIDAR). Our sound detection and localization can use that to **prune improbabilities** (like classifying a vehicle as a child).

References

[1] J. Valin, F. Michaud, J. Rouat, D. Letourneau, Robust sound source localization using a microphone array on a mobile robot, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 2, 2003, pp. 1228–1233.

[2] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 5, pp. 971–982, May 2013.