Acoustic Localization and Classification of Sound Sources for Autonomous Vehicles

Abhishek Das
ECE Department, CMU
addas@andrew.cmu.edu

Soham Shailesh Deshmukh
ECE Department, CMU
sdeshmuk@andrew.cmu.edu

Swapnil Das
ECE Department, CMU
swapnild@andrew.cmu.edu

Revanth Banala Nithyanandam
ECE Department, CMU
rbanalan@andrew.cmu.edu

1 Introduction

The audio-based perception system of an autonomous vehicle (AV) is as important as the vision-based perception for discerning its environment. This allows an AV to classify the type of audio signal and locate the sound source by auditory input alone. With these capabilities, it can complement existing vision-based perception and pave the way for multi-modal operation where differing sensors such as LIDAR/RADAR work together alongside sound to paint a fuller image of the surroundings.

This problem is an important want to tackle - we want that autonomous vehicles be resilient and able to handle a dynamic environment. Just as human drivers use both sight and sound to gauge the surroundings, we intend the same for AVs. For instance, an AV that relies on LIDAR to detect a nearby vehicle can inform itself to make a better decision if it could hear a siren; that way, it would know that the vehicle is special (be it a fire truck, police car or ambulance) and would aim to give way. Audio detection can help detect the presence of nearby trains far more reliably than LIDAR or rail crossing signs that may or may not exist. These are just some real-world examples where having audio as a new sense can bolster the reliability and performance of AVs.

Our project seeks to take steps toward tackling this impetus, applying machine learning techniques to help classify the sounds of objects in the vicinity of a car, as well as localize them.

2 Evaluation of Baseline Work

Studies so far have shown that the research on audio-based perceptual systems is still a novel part of advanced perception in the context of autonomous vehicles; this establishes the need for systems that automatically analyze complex and rich information taken from different sensors in order to obtain refined information on the sensed environment and the activities being carried out within them.

Past studies have been done on localizing sound sources in an indoor environment using deep learning and general signal processing techniques for robots. For our case, we will seek methods that avoid the use of deep learning techniques. We intend to build on the indoor acoustic localization scaling it to an outdoor scenario with an indeterminate number of sound sources.

To this effect studies have also been done in source separation given multiple channels of signals, which is relevant to us because we need multiple channels in order to perform a satisfactory localization.

3 Methods

Because we can be potentially surrounded by an indeterminate number of sounds in our environment, our project seeks to take steps towards tackling this by applying machine learning techniques to help separate, classify and localize sounds of objects in a car's vicinity. The overall system and flow is depicted in the diagram below. In the 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.

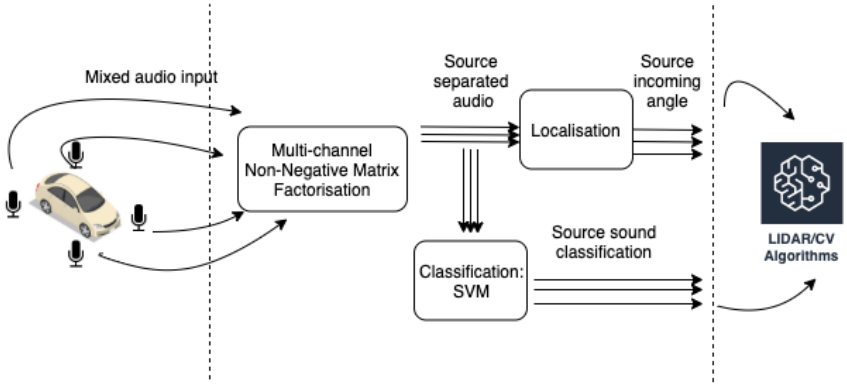

Figure 1: Overall system diagram

3.1 System Setup

In order to do any experiments, we need to mimic the set up of a car. We do this by capturing microphone inputs with an Alesis IO4 with four channels. With each of the four mics set at a different corner of a table, we can record audio and test with it as needed.

We need four channels because sound is read differently by each mic which is used for localization; each mic encodes unique spatial and temporal information. The fact that we don't have a moving vehicle doesn't matter - this is because our localization module determines the positions of sound sources relative to the vehicle.

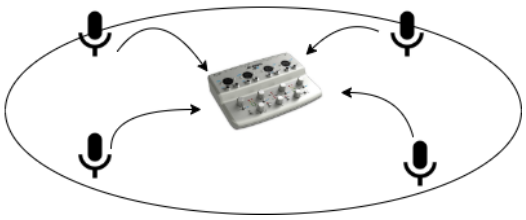

Figure 2: System setup for recording multichannel audio

3.2 Source Separation

The first step is the separation of the sounds in the environment. In our project, we look at Non-negative Matrix Factorization (NMF) as our desired choice of technique. The underlying motivation behind this option was so that individual bases and weights would be more easily interpretable and analyzable for future use; the use of other methods like ICA would involve negative bases and weights which have no physical meaning. However, given that we are using multiple audio channels (for proper localisation), we use a variant of NMF called Multi-channel NMF (MNMF)

MNMF varies from the traditional algorithm as it encodes signal magnitude and phase differences between channels in a matrix H (which we can also apply in localisation). As a result, there are

different update rules that we have to account for. Below are the update rules that take into account this information (the terminology denotes T as the basis matrix, and V as the weight matrix).

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_j v_{kj} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{ik})}{\sum_j v_{kj} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{ik})}}$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_i t_{ik} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{ik})}{\sum_i t_{ik} \mathrm{tr}(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{ik})}}.$$

$$\mathsf{H}_{ik} \mathsf{A} \mathsf{H}_{ik} = \mathsf{B}$$

$$\mathsf{A} = \sum_j v_{kj} \hat{\mathsf{X}}_{ij}^{-1}, \quad \mathsf{B} = \mathsf{H}'_{ik} \left(\sum_j v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \right) \mathsf{H}'_{ik}$$

where H'_{ik} is the target matrix before the update.

However, MNMF has a long running time, which deters real-time usage. (The reason for this long running time is because each iteration also involves the H update rule, which is the computational bottleneck.) While the results are promising, we need to combat this; therefore, 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 - we use the same update rules as above but only for H, keeping T and V constant. Thus, we learn an approximated H we call H_a. The intuition here being the B that construct the music should remain the same, and hence the algorithm can be warm-started using the B from traditional NMF. The speed is improved drastically since the bases and weights are already learned, minimizing the number of iterations needed to obtain an approximate H, and the information is preserved as best as possible in this approximation since we use data from all the microphones.

3.3 Acoustic Source Localization

The preliminary implementation of the sound source localization was done using a self-generated phase-variant forms of acoustic signals from the Urban Sound Dataset. With the above setup, each microphone "perceives" the same signal with a certain time delay corresponding to the phase delay in the signal.

The localization model is implemented based on the concept of Time-Difference-of-Arrival(TDOA), by finding the cross-correlation between the phase variant signals in the frequency domain. When the correlation lag is equal to the offset between the signal being perceived by both microphones, then the correlation is maximum. Due to the noise and low-pass signals, the correlation peaks are widened, hence we use the "whitening" concept prior to finding the correlation matrix, which is given by the new "whitened cross-correlation matrix":

$$R_{ij}^{(w)}(\tau) = \sum_{k=0}^{N-1} \frac{X_i(k) X_j(k)^*}{|X_i(k)| |X_j(k)|} e^{i 2\pi k \tau / N}$$

On using the above whitened form, we eliminate the wider frequency components and associate the phase to the correlation between the signals by adding the spectral weights to find the peaks. Based on far field method, we localize the source based on the difference in the phase delay of the same signal captured by multiple microphone stored in various arrays. On finding the correlation peaks, the direction and distance of the source will be estimated as shown below, using the cosine law.[12]

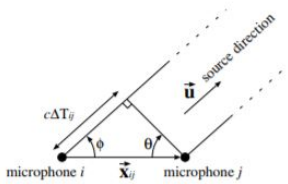

Figure 1. The estimation of the 2-D distance and locating the source direction from the signals captured by the model.

$$\cos\phi = \sin\theta = \frac{c\Delta T_{ij}}{\|\vec{\mathbf{x}}_{ij}\|}$$

$$u(x_j - x_i) + v(y_j - y_i) + w(z_j - z_i) = c\Delta T_{ij}$$

$$\begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) & (z_2 - z_1) \\ (x_3 - x_1) & (y_2 - y_1) & (z_3 - z_1) \\ \vdots & \vdots & \vdots \\ (x_N - x_1) & (y_N - y_1) & (z_N - z_1) \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} c\Delta T_{12} \\ c\Delta T_{13} \\ \vdots \\ c\Delta T_{1N} \end{bmatrix}$$

where u = (u, v, w) and x_{ij} = (x_j x_i, y_j y_i, z_j - z_i), the position of microphone i being (x_i, y_i, z_i,). Using the above mathematical localization model, we estimate the X, Y and Z components to determine the direction vector with respect to the autonomous vehicle at the origin.

3.4 Sound Source Classification

For doing sound source classification, time series and amplitude based where extracted from each snippet of audio. This divides the input signal into short-term windows (frames) and then computed 34 features for each frame. This process leads to a sequence of short-term feature vectors for the whole signal. The features extracted are:

Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral spread, Spectral Entropy, Spectral Flux, Spectral roll-off, MFCC, Chroma vector, Chroma deviation

After features from each seperated sound are extracted, we used different models namely Support Vector Machines with Linear and Gaussian kernels, Decision Tree, K-Nearest Neighbours, and Logistic Regression. The training of these models was done using three classes of interest (sirens, dogs barking and children playing on the street) from the UrbanSound dataset. A reason of picking these three specific classes was that they are pertinent for an autonomous vehicles to make certain driving decisions and at the same time they encompass high, medium and low frequency sounds which ensures better generalization for the model on the test set.

4 Datasets

As far as our datasets, we have used the UrbanSound dataset for our sounds that we trained our classification with. This dataset has had pros and cons both - while the dataset itself has high-quality samples of sounds (like cars, sirens, street music etc.) However, the actual number of samples is low, leading to potential consequences for the classification step.

Another stage of our dataset collection is in our actual testing of localisation and the whole pipeline, which was achieved by using our Alesis IO4 and recording sounds around it, moving audible speakers to mimic the relative velocities of nearby vehicles. While this let us test our data, the actual collection of it was time-consuming and inefficient.

5 Results and Discussion

On testing the system with the mixed-audio inputs of sound signals belonging to three major classes - "sirens", "children" and "dogs", the system performs well in separating the sound sources into the component signals and identifying the signal class. Once the phase-variant signals are fed into the localization algorithm, we notice that the system is accurate in localizing the sound source in terms of the direction.

5.1 Source Separation

There are two factors to consider in the evaluation of source separation - the quality of the obtained H, and the quality of the separated audio. Below we provide first graphs that show the quality of our approximation of H. We do this by comparing the diagonal and off-diagonal elements that capture information about power and phase respectively. We can do this by finding the difference of norms of diagonals and off-diagonal elements separately, as shown below. We observe how the difference is minimized in the first iterations, hence establishing that our choice for speedup is justified in only considering the first couple iterations.

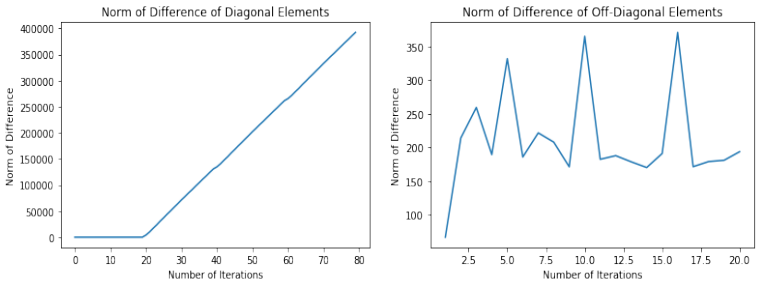

Figure 3: Norm Difference of diagonal power elements (left) and off-diagonal phase elements (right)

Below, we also provide data that shows the quality of our sound approximations and reconstructions with our separation algorithm as below. As we can see, we can note the visible similarities in the shape of the spectrogram of the original and the reconstructed audio, thus confirming the confidence of our actual separation.

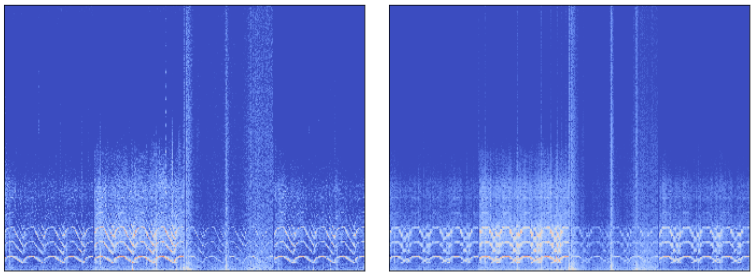

Figure 4: Spectrogram of police siren with noise (left) and reconstructed spectrogram of police siren when k = 5 and iterations = 50 (right)

5.2 Acoustic Source Localization

The localization algorithm finds the cross-correlation matrix between the each set of microphones and "whitens" the matrix. Once "whitened", spectral weights are added in order to suppress the noise and elevate the peaks of the signal. The whitened cross-correlation matrix and the original spectrogram have been showed in the below figure.

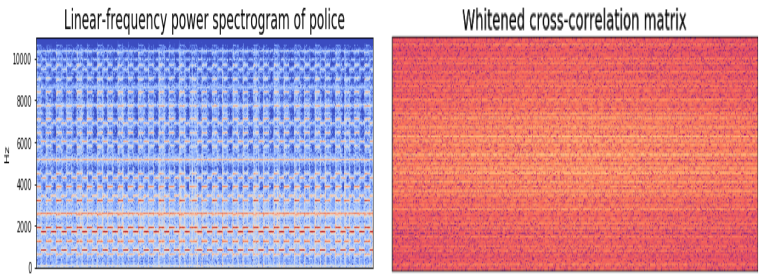

Figure 5: Spectrogram (left) and Whitened cross-correlation matrix of police siren (right)

On finding the maximum peaks and passing them to solve for the X-Y-Z coordinates of the sound source with respect to the microphone setup as the origin, we estimated the direction of the sound source. In reality, the direction of the sound source and coupling it with the actual distance calculated by a LIDAR in the existing autonomous vehicles, we can localize the sound source more accurately.

On finding the direction vectors for a constantly moving sound source with respect to the setup, we can accurately model the motion of the point-clouds and hence localize the motion of the sound source. This will able to keep the autonomous vehicle "well-informed" about the approaching vehicle and act accordingly. As we will see in the challenges that we faced in localization, our approach of quantifying localisation results was not up to the mark - we have considered solutions for future exploration to ground the results we have obtained better.

5.3 Sound Source Classification

The accuracy of different classification models as well as the precision of those models for each class is depicted in the plots below.

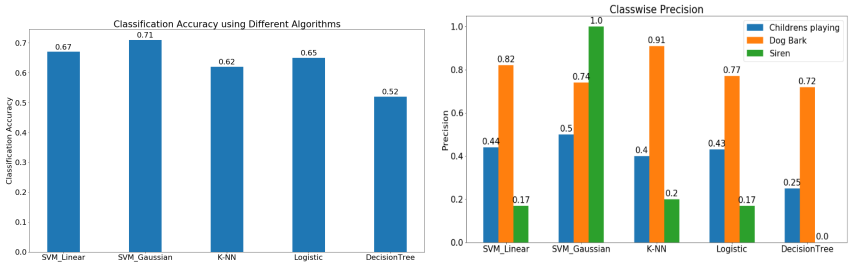

Figure 6: Accuracy (left) and Class-wise precision (right) plot

As discussed before in the challenges faced with the UrbanSound dataset, the low number of samples meant that accuracy would suffer as a result. We see the best performance in the SVM-Gaussian as above. In an ideal real-world integration with an AV, vision data can be used as a prior to bolster the audio predictions that this system makes, allowing improbabilities to be pruned faster (for instance, classifying a vehicle as a child).

6 Challenges and Scope of Future Work

One of the major challenges faced during the project implementation was quantifying the localization results as the manual measurement is error-prone and cannot give accurate results. One possible way to solve this problem is to run a robotic car (with the microphone setup) on a grid-layout and playing the mixed audio inputs at known locations on the grid. Using this technique, we can accurately estimate the direction and compare it with actual coordinates to quantify the error in localization. This poses as a potential research topic in the field of autonomous vehicles.

In terms of improvements that we could make to our source separation, there are several other approaches that we could consider to provide optimizations while still retaining the robustness of the original MNMF - a solution would be frequency binning, reducing the dimensionality of our input by focusing on relevant frequency bins. This way, mathematical calculations would be faster (though the determination of the H matrix would yet remain a bottleneck). NMF also scrubs away phase information inherently, so that may hamper with our intent of using H; other methods that preserve intuition and real-world sense can be considered that still preserve that information. MNMF also reportedly has the restriction of being able to reliably determine at most 3 sounds in the environment - while our approximation algorithm relaxes those constraints and more sounds should be able to be detected in theory, there are also areas to consider whether the limit can be pushed upwards.

A potential idea for the future (for our live data collection) is to make some sort of simulation for sound samples - by setting a particular model of sound propagation, we can feed transformed audio into each channel and bypass having to use the microphones for any testing. This can prove to speed up the pipeline of work - however, it will require dedicated and proper parameterizing of the mics to make sure that our transformations are accurate.

7 Conclusion

With the rise of the era of autonomous vehicles, sensing and perception is quite critical for their robustness. Audio perception builds an additional layer to the existing perception systems making the autonomous more resilient and fault-tolerant by complementing the vision based perception. The localization algorithm built during our project implementation, once tested with a vision based-system should be able to quantify an approaching acoustic source, hence enhancing the performance of autonomous vehicles. With separation, multiple sounds can be separately discerned and subsequently processed, giving information to the AV. Our project has therefore taken concrete steps towards achieving our goal laid out in the introduction.

References

[1] Rascon, C., Meza, I. Localization of sound sources in robotics: A review (Open Access) (2017) Robotics and Autonomous Systems, 96, pp. 184-210. doi: 10.1016/j.robot.2017.07.011 https://www.sciencedirect.com/science/article/pii/S0921889016304742

[2] Argentieri S., Danès P., Souères P. A survey on sound source localization in robotics: From binaural to array processing methods Comput. Speech Lang., 34 (1) (2015), pp. 87-112

[3] Xiaofei L., Hong L. A survey of sound source localization for robot audition CAAI Trans. Intell. Syst., 7 (1) (2012), pp. 9-20

[4] Irie R.E. Multimodal sensory integration for localization in a humanoid robot Proceedings of IJCAI Workshop on Computational Auditory Scene Analysis, (CASA), Morgan Kaufmann Publishers, Inc. (1997), pp. 54-58

[5] Takanishi A., Masukawa S., Mori Y., Ogawa T. Development of an anthropomorphic auditory robot that localizes a sound direction Bull. Centre Inform., 20 (1995), pp. 24-32

[6] Kumon M., Shimoda T., Kohzawa R., Mizumoto I., Iwai Z. Audio servo for robotic systems with pinnae Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), IEEE (2005), pp. 1881-1886

[7] Keyrouz F., Maier W., Diepold K. A novel humanoid binaural 3d sound localization and separation algorithm Proceedings of IEEE-RAS International Conference on Humanoid Robots, IEEE (2006), pp. 296-301

[8] V.M. Trifa, A. Koene, J. Morén, G. Cheng, Real-time acoustic source localization in noisy environments for human–robot multimodal interaction, in: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN, 2007, pp. 393–398.

[9] Kim H.-D., Komatani K., Ogata T., Okuno H.G. Human tracking system integrating sound and face localization using an expectation-maximization algorithm in real environments Adv. Robot., 23 (6) (2009), pp. 629-653

[10] A. Portello, P. Danès, S. Argentieri, Acoustic models and kalman filtering strategies for active binaural sound localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 137–142.

[11] Chen J.C., Yao K., Hudson R.E. Acoustic source localization and beamforming: Theory and practice EURASIP J. Adv. Signal Process., 2003 (4) (2003), p. 926837

[12] 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.

[13] M. Omologo, P. Svaizer, Acoustic event localization using a crosspower-spectrum phase based technique, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Vol. 2, 1994, pp. II/273–II/276.

[14] Markovic I., Petrovic I. Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering Robot. Auton. Syst., 58 (11) (2010), pp. 1185-1196

[15] U.-H. Kim, T. Mizumoto, T. Ogata, H. Okuno, Improvement of speaker localization by considering multipath interference of sound wave for binaural robot audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 2910–2915.

[16] Takeda R., Komatani K. Discriminative multiple sound source localization based on deep neural networks using independent location model Proceedings of the IEEE Spoken Language Technology Workshop, (SLT), IEEE (2016), pp. 603-609

[17] Murray J.C., Erwin H.R., Wermter S. Robotic sound-source localization architecture using cross-correlation and recurrent neural networks Neural Netw., 22 (2) (2009), pp. 173-189

[18] Nakamura K., Nakadai K., Okuno H.G. A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition Adv. Robot., 27 (12) (2013), pp. 933-945

[19] Tanabe R., Sasaki Y., Takemura H. Probabilistic 3d sound source mapping system based on monte carlo localization using microphone array and lidar J. Robot. Mechatronics, 29 (1) (2017), pp. 94-104

[20] Tourbabin V., Rafaely B. Direction of arrival estimation using microphone array processing for moving humanoid robots IEEE/ACM Trans. Audio Speech Lang. Process., 23 (11) (2015), pp. 2046-2058

[21] 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.