Poster Abstract: Improving Acoustic Detection and Classification in Mobile and Embedded Platforms

Stephen Xia
Columbia University
New York, New York, USA
stephen.xia@columbia.edu

Xiaofan Jiang
Columbia University
New York, New York, USA
jiang@ee.columbia.edu

ABSTRACT

Sound detection and classification are critical in many acoustic-based applications. Existing works generally focus on discovering new features and classifiers to improve detection. However, in many scenarios the presence of other sounds may hinder the performance of these sound classifiers. In this work, we take a sound filtering and enhancement approach to improve sound detection for mobile and embedded applications, regardless of the type of detector used.

KEYWORDS

sound source separation, adaptive beamforming, acoustic detection

ACM Reference Format:
Stephen Xia and Xiaofan Jiang. 2021. Poster Abstract: Improving Acoustic Detection and Classification in Mobile and Embedded Platforms. In *The 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021) (IPSN '21), May 18–21, 2021, Nashville, TN, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3412382.3458784

1 INTRODUCTION

Sounds provide a large amount of information and helps us make informed decisions in daily lives. For instance, hearing your child crying may be a sign that he needs parental attention or nourishment. Hearing your dog bark may be a sign that an intruder is stepping onto your property. Sometimes, we just listen to music for entertainment. Many works based on acoustics have been proposed to solve pressing problems. For instance, [1–3] propose a set of acoustic wearables to detect, localize, and warn users of potentially dangerous oncoming vehicles in efforts to reduce vehicle and pedestrian accidents. There have also been numerous works [4] and smartphone applications [5] developed that use the microphone on mobile devices to monitor sleep quality.

However, in many real-world scenarios, sounds that are of interest may be overpowered by other sounds in the environment. For instance, a noisy construction site may obscure the sound of an approaching vehicle for a pedestrian passing by. In such scenarios, acoustic detectors may perform worse due to heavy noise.

IPSN '21, May 18–21, 2021, Nashville, TN, USA

ACM ISBN 978-1-4503-8098-0/21/05…$15.00
https://doi.org/10.1145/3412382.3458784

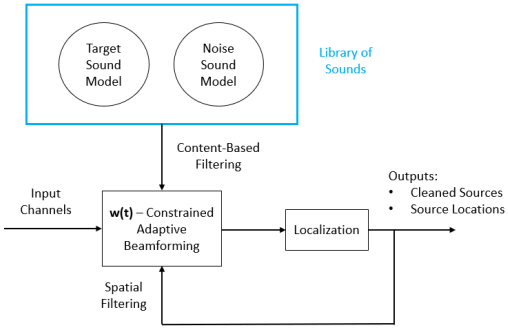

Figure 1: System Architecture.

In this poster, we present a general acoustic framework that combines adaptive beamforming and learned models of specific sounds to filter or enhance them to improve sound detection.

2 SYSTEM IMPLEMENTATION

2.1 Adaptive Beamforming with Content-Based Constraints

We propose an adaptive beamforming method that combines content-based constraints based off of learned models of sounds that users can select to enhance or filter out depending on the application requirements. This framework can accommodate a wide range of mobile and embedded applications that may have varying numbers of acoustic sensors. These ideas are shown in Equations 1 and 2.

$$\arg\min_{w} w^* R w \quad (1)$$
$$w^* d = 1$$

$$P_n(w^* R w) < \alpha \quad (2)$$
$$P_e(w^* R w) > \beta$$

Equation 1 shows the problem set up for the linearly constrained minimum variance adaptive beamformer (LCMV) [6]. R is defined as the spatial correlation of the input signal, x ($R = E[xx^*]$). $E[\cdot]$ refers to the expected value, and the $*$ operator refers to the conjugate transpose. w refers to the filter coefficients that the adaptive beamformer learns to enhance signals arriving at the microphone array from the direction specified by the steering vector d.

Many applications may have information or models of sounds that they wish to detect or reject. To incorporate this information into our framework, we introduce the constraints presented in Equation 2 into the adaptive beamforming framework presented

in Equation 1. We refer to a sound of class n as a sound the application wishes to reject and a sound of class e as a sound that the application is interested in enhancing or detecting. The idea behind these constraints is to adapt our filter coefficients w to reduce the probability that the filtered signal, $P_n(w^*Rw)$, is of the undesired class n and to improve the detection rate or probability, $P_e(w^*Rw)$, that the filtered signal is recognized as an instance of the desired class e. The final filter update equations can be obtained by solving the optimization problem using the method of Lagrange multipliers similar to what is used in the original LCMV beamformer [6].

Two problems remain that have yet to be addressed. First, beamforming requires a direction to "steer" the beam towards. To address this issue, we integrate the sound source localization scheme presented in [7] to detect sources present in the environment and localize them. The steering vector d is then updated to "steer" the beam to the direction of localized sources.

The second issue is modeling the sounds we wish to filter out or enhance ($P_n(w^*Rw)$ and $P_e(w^*Rw)$). We model our sound models using Gaussian Mixture Models (GMM). This is because Gaussians and mixture models are commonly used to model acoustic sounds and computing gradients for solving the adaptive beamformer with content-based constraints presented in Equations 1 and 2 is straightforward and relatively computationally inexpensive.

2.2 System Architecture

Figure 1 shows the architecture of our system. The central piece is the adaptive beamforming block, introduced in Section 2.1. Once the audio channels pass through the beamforming block, our system performs sound source localization. The source locations and filtered sources are fed back into the beamforming block in an adaptive manner to continue the filtering process for the next window.

The library of sounds block, highlighted in blue, contains a set of sound models that users can choose to enhance or filter out in the beamforming block, depending on the application at hand.

3 PRELIMINARY RESULTS

We explored two scenarios where our system may be useful. First, is the case where we wish to detect the crying sounds of a baby in presence of loud construction sounds. This scenario may be part of an application that determines when the child needs care, but may be difficult to detect due loud city sounds. The second scenario is the case where we wish to detect the sounds of a piano playing in presence of speech and babble. This scenario may occur in situations where the music is being played at a large social gathering and attendees may wish to listen to the background music.

We collected and recorded 10 minutes of clean audio for each of the four classes mentioned. For both scenarios, we recorded 10 minutes of mixtures. In total, we recorded 60 minutes of audio (40 minutes in total of clean sounds and 20 minutes of mixtures). All sounds were recorded by playing clips collected from the Google Audioset dataset [8] through speakers pointed at a 6-microphone uniform circular array to ensure that the sounds are mixed in the real-world. We used 20% of our data as testing data to compare the performance of a sound detector on the raw unfiltered signal vs the sound signal obtained after processing the raw audio through our proposed architecture. The other 80% of data was used to train our sound models used for filtering and a Random Forest sound detector to perform detection on the filtered and unfiltered signals. The confusion matrix results are shown in Table 1.

Table 1: Performance metrics of two scenarios: 1. crying detection in presence of construction tool sounds and 2. piano detection in presence of speech. Detection performance after filtering through our proposed pipeline is in bold.

	True Pos.	True Neg.	False Pos.	False Neg.	F-1
Crying Nonfiltered	80%	98%	2%	20%	0.88
Crying Filtered	**83%**	**98%**	**2%**	**17%**	**0.90**
Piano Nonfiltered	86%	98%	2%	14%	0.90
Piano Filtered	**90%**	**98%**	**2%**	**10%**	**0.92**

We see in both scenarios that filtering signals through our pipeline improves the true positive rate of crying and piano (the target signals). This in turn also improved the F-1 score in both scenarios, showing that we can obtain better detection accuracy by preprocessing acoustic signals through our proposed system.

4 CONCLUSION

In this poster, we presented an acoustic filtering framework for improving sound detection and classification for mobile and embedded platforms. Our system utilizes a novel adaptive beamforming method that is constrained by sound models of noises and target sounds a user or application can choose to reduce or enhance.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under Grant Numbers CNS-1704899, CNS-1815274, CNS-11943396, and CNS-1837022. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Columbia University, NSF, or the U.S. Government or any of its agencies.

REFERENCES

[1] Stephen Xia, Daniel de Godoy Peixoto, Bashima Islam, Md Tamzeed Islam, Shahriar Nirjon, Peter R Kinget, and Xiaofan Jiang. Improving pedestrian safety in cities using intelligent wearable systems. *IEEE Internet of Things Journal*, 6(5):7497–7514, 2019.

[2] Daniel de Godoy, Bashima Islam, Stephen Xia, Md Tamzeed Islam, Rishikanth Chandrasekaran, Yen-Chun Chen, Shahriar Nirjon, Peter R Kinget, and Xiaofan Jiang. Paws: A wearable acoustic system for pedestrian safety. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 237–248. IEEE, 2018.

[3] Stephen Xia, Daniel de Godoy, Bashima Islam, Md Tamzeed Islam, Shahriar Nirjon, Peter R Kinget, and Xiaofan Jiang. A smartphone-based system for improving pedestrian safety. In *2018 IEEE Vehicular Networking Conference (VNC)*, pages 1–2. IEEE, 2018.

[4] Stephen Xia and Xiaofan Jiang. Pams: Improving privacy in audio-based mobile systems. In *Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*, pages 41–47, 2020.

[5] Urbandroid. Sleep as android (version 20200806), 2010. [Mobile app]. Retrieved from https://play.google.com/.

[6] O. L. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

[7] Anastasios Alexandridis, Anthony Griffin, and Athanasios Mouchtaris. Capturing and reproducing spatial audio based on a circular microphone array. *JECE*, 2013, January 2013.

[8] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.