Major Changes

While, there have been no major changes in my project since my proposal, a cursory review of the literature has revealed that each of the original-proposed milestones might take longer than anticipated. Hence, I have modified my timeline(see below) to make it more realistic.

In terms of implementation, I have decided to work with PyTorch and use reinforcement learning in order to identify sources and separate the different sources of sound.

Accomplishments so far

1. **Learned Pytorch:** While I have worked with Tensorflow and Keras for other Machine learning projects, I have never used PyTorch before. However, my team (the Multimodal Machine Learning group at CMU) primarily uses PyTorch for their Machine Learning Models. As a result, I used this time to learn PyTorch and implemented simple algorithms.

2. **Identifying Current State of the Art:** I was unable to identify any researches that use source separation using ego-centric videos or find any data sets that apply to this case perfectly. However, I have been able to identify several important research papers that might be helpful to the current project. I find one of these particularly interesting (Co-separation of visual objects)[1] and would be re-implementing the algorithm described in the paper as a baseline model. This algorithm identifies object-level sounds from unlabeled multi-source videos and claims to have outperformed its predecessors in order to do so. The main contribution of the paper is using object-level sounds from unlabeled multi-source videos and doing so with high accuracy even when the individual sound isn’t heard.

3. **Data set creation:** After talking to the Multimodal Machine Learning group I was suggested to work on traditional data sets and try to develop a good algorithm for identifying the sound of the object as a milestone. For this, I helped in the creation of a audio-visual data set using the Google Audiovisual data set. A subset of these videos were labeled for the atomic sounds (100 videos per each category for 100 categories). I plan to use this data set and leverage the co-separation technique for semi-supervised learning of the model as a starting point. The initial model I make will be identifying different objects in the frame, and trying to separate the sound of the object 'selected'(using a mouse) on this data set.

Meeting Milestone

I was able to achieve the milestone for dataset creation. However, this process, learning pytorch and literature review took longer than expected so I wasn’t able to create a baseline
model.

**Surprises**

I had expected to find an ego-centric data set that, though, unlabelled might help in my current project. However, after searching online, I was unable to find any ego-centric data set that fits my needs. Most of the data sets present online do not have sound and therefore won’t be useful in the project. As a result, I planned to work on a traditional data set in the beginning. Once this part of the project is completed, it is expected to carry-over well to an egocentric setting well. I will then move on to creation of an ego-centric dataset from scratch, potentially by outsourcing the video collection process using the 'Spring Opportunities' document.

Since I have no background in the creation of hardware and the machine learning component is expected to take more time than originally anticipated, I have asked a master’s student working with me in the Audio-visual team to help me with this part of the project.

**Revision to Milestones**

1. **January 27th:**
   - Starting the implementation of the Co-separation algorithm on PyTorch and brainstorming ideas for other machine learning models

2. **February 10th:**
   - Identification of a few best models and trying to use semi-supervised learning using these algorithms. Parallely start the process of data collection for creation of a basic headphone with two cameras.

3. **February 24th:**
   - Further tuning the machine learning model, reporting observations. Start data-collection using headphone

4. **March 16th:**
   - Improvising the speed/identifying pre-processing techniques that might help in improved accuracy.

5. **March 30th:**
   - Developing a machine learning model that identifies what the person is pointing at. If the data-collection process is not complete yet, I will start the development of this model on an existing data set (After a cursory review of literature, several such data sets were found)
6. *April 14th:*
   Adapting the current model to the new data set.

7. *April 27th:*
   Further working on the final prototype.

**Resources needed**

I have obtained access to the multimodal machine learning cluster so that I can work on my machine learning algorithm. In addition, I have installed all the required software (i.e., PyTorch).

I still need to acquire the hardware for the headphone that will be ordered online.

**Citations**