# Keyword Spotting Dataset Creation

As we just learned, there is a lot of work that goes into designing a good dataset for keyword spotting. And this is true for all applications of TinyML. This is because the requirements for different applications (and thus the data that needs to be collected) can vary drastically even for the same kinds of inputs. In fact, even for the same keyword deployed into different environments!

This makes it impossible to come up with hard and fast rules for how to design a good dataset. But/and here are a couple of things you might want to consider when designing an audio data collection scheme.

## Who are the anticipated end users?

It is important to understand who the end users might be as they may engage with the application differently and may present different kinds of challenges for how you need to design the application and what data you will need to collect. For example:

- What languages will they speak?
- What accents will they have?
- Will they use slang or formal diction?
- Will the users speak clearly?

## In what environments will the users employ the application?

Imagine a Keyword Spotting device. Is it placed in a quiet room where it can hear you clearly? Or is it placed right next to your TV in the living room? The environment can greatly affect the quality and type of audio that can be collected, as well as the amount of additional sounds that must be accounted for. For example:

- How much background noise do we expect?
- Who/how many people will be talking in the environment?
- How far will the users be from the device and sources of noise?
- Will the user be stressed vs. calm vs. panicked?
- Will the user use different volumes of voice (whispered/normal/loud/shouted)?
- How likely is it that these keywords may be triggered unintentionally during conversation?

## What are the goals for using the application? How will this impact the requirements of the ML model's performance?

Depending on the goals of the application, model errors can either have very little consequences, or they can be catastrophic. This can directly impact how robust of a data collection scheme you need to implement. For example consider the difference between these three scenarios:

- The user is trying to turn the thermostat up/down
- The user is trying to arm/disarm a security system
- The user is trying to play their favorite playlist from a smart speaker

## Given all of these previously mentioned factors here a short checklist to consider when implementing your final data collection scheme:

- What custom keywords will you select?
- What background noise samples do you need to collect to augment your dataset so that the trained model is robust?
- What other words do you need to collect to ensure the model learns the difference between them and your keywords? It is important to collect other words so that the model can learn your particular word and not simply the general sounds of humans talking!
- How much data will you need to collect? We all know more is better (usually), but you also live in the real world with time constraints so how much do you think will be enough?
- In what environments will you collect these samples?